Automatisierte Erkennung und Evaluation von therapeutischen Übungen für Personen mit Mimikdysfunktionen

Dissertation

Zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Informatik und Automatisierung
der Technischen Universität Ilmenau

von

Dipl.-Ing. Cornelia Dittmar

Gutachter:

1. Prof. Dr.-Ing. Horst-Michael Groß
2. Prof. Dr.-Ing. Joachim Denzler
3. Prof. Dr.-Ing. Dietrich Paulus

Tag der Einreichung: 03.06.2016
Tag der wissenschaftlichen Aussprache: 07.12.2016

urn:nbn:de:gbv:ilm1-2016000698

Kurzbeschreibung

In dieser Arbeit wird ein flexibles, kamerabasiertes Trainingssystem zur Rehabilitation von Gesichtslähmungen (Fazialisparesen) und anderen Mimikdysfunktionen vorgestellt. Das System unterstützt das selbstständige Training des Patienten, indem es die Durchführung von insgesamt zwölf Fazialisübungen automatisch bewertet und mehrstufiges Feedback an den Anwender vermittelt. Es eignet sich somit für einen begleitenden Einsatz zu den regulären Übungseinheiten, welche von einem Logopäden oder Sprechwissenschaftler angeleitet werden.

Während Ansätze zur automatisierten Diagnose und Gradierung von Fazialisparesen in der Literatur vergleichsweise verbreitet sind, finden sich gegenwärtig nur vereinzelt Konzepte für therapiebegleitende Trainingsanwendungen. Die diesen Anwendungen zu Grunde liegenden Algorithmen sind zudem auf einzelne Fazialisübungen spezialisiert und daher, anders als das in dieser Arbeit vorgestellte System, nicht ohne Mehraufwand auf weitere Übungen übertragbar.

Die Beiträge der vorliegenden Arbeit umfassen die wesentlichen Komponenten der technischen Gesamtarchitektur des Trainingssystems. Der methodische und experimentelle Fokus der Ausarbeitung liegt dabei vor allem auf der Merkmalsextraktion, sowie der Ableitung des Feedbacks aus den extrahierten Merkmalsdeskriptoren. Eine wesentliche Neuheit gegenüber dem Stand der Technik besteht in der Möglichkeit, das Trainingssystem flexibel um zusätzliche Fazialisübungen zu ergänzen und sowohl globales als auch regionenbezogenes Feedback bereitzustellen.

Die dafür ausgewählten Verfahren basieren vorwiegend auf der Verarbeitung von 3D-Kameradaten und umfassen die Extraktion von Punktsignaturen, Histogrammen orientierter Normalenvektoren, sowie von Krümmungs-, Distanz- und Winkelmerkmalsdeskriptoren. Die Feedbackermittlung stützt sich auf den Einsatz von Random-Forests und den aus diesen ableitbaren paarweisen Ähnlichkeiten. Letztere stellen Schätzwerte für die merkmalsbezogene Übereinstimmung zwischen der vom Patienten ausgeführten Übung und den Modelldurchführungen in den Trainingsdaten dar.

Abstract

This thesis presents an automated, camera-based training system employable for the therapy of facial paralysis and related muscle dysfunctions. The proposed system aims to support patients in conducting twelve different facial exercises by providing automatically generated feedback. Thus, it is suited to supplement individual exercise sessions that are not supervised by a therapist.

Automated grading and diagnosis systems for facial paralysis are a prominent topic in the literature on clinical image processing. In contrast, only few papers deal with the development of automated training systems for facial muscle re-education. Furthermore, the underlying algorithms are typically specialized for particular facial exercises and difficult to adapt to additional requirements.

The contributions of this thesis comprise the main components of the system architecture with a methodical and technical emphasis on feature extraction algorithms and feedback estimation methods. Regarding the state-of-the-art, the major novelty is embodied in the possibility to easily extend the system to additional exercises and in the derivation of global and local feedback.

The selected approaches rely on processing of 3D-camera data and include the extraction of point signatures, histograms of oriented normal vectors, curvatures, distance, and angle features. The feedback generation is based on random forest classifiers and proximities derived from trained forests. These proximities provide an estimate of similarity between the patient sample and training data samples.

Danksagung

An dieser Stelle möchte ich allen danken, die mich auf dem Weg zu dieser Arbeit begleitet und zu ihrer erfolgreichen Umsetzung beigetragen haben.
Mein erster Dank gilt meinem Betreuer Prof. Dr. Horst-Michael Groß, welcher mich durch zahlreiche gemeinsame Diskussionen und wertvolle fachliche Anregungen wesentlich bei der Entstehung dieser Arbeit unterstützt hat.
Besonderer Dank gilt auch Prof. Dr. Joachim Denzler für seine inhaltlichen Impulse und Beiträge, die einen wichtigen Grundstein zu dieser Arbeit gelegt haben.
Ich danke allen Mitarbeitern des Fachgebiets für Neuroinformatik und Kognitive Robotik (TU Ilmenau) und des Lehrstuhls für Digitale Bildverarbeitung (FSU Jena) für die kollegiale Arbeitsatmosphäre und die schöne gemeinsame Zeit.
Zu erwähnen sind auch alle Studenten, welche im Rahmen von studentischen Arbeiten oder Hiwi-Jobs wertvolle Zuarbeit geleistet haben. Hervorheben möchte ich Birant Sibel Olgay, unter anderem für die Unterstützung bei der Datenaufnahme und dem Labeln von über 1400 Bildern während der Feiertage.
Herzlicher Dank geht auch an Prof. Dr. med. Gustav Pfeiffer, Eva Schillikowski und Anke Arnold von der Fachklinik Bad Liebenstein, sowie Irina Stangenberger von der Logopädiepraxis Stangenberger für ihre fachliche Beratung. Besondere Hervorhebung verdienen die Patienten der Fachklinik Bad Liebenstein, die einen Einblick in ihren Therapiealltag gestattet und sich für die Aufnahme eines Datensatzes zur Verfügung gestellt haben.
Für die finanzielle Unterstützung danke ich den Organisatoren und Mitgliedern der Graduiertenschule Bildverarbeitung und Bildinterpretation, sowie dem Gleichstellungsrat der TU Ilmenau (insbesondere Dr. Karin Bieske).
Zu guter Letzt möchte ich meiner Familie, vor allem meinen Eltern und meinem Ehemann, danken.

Inhaltsverzeichnis

1. Einleitung

Die vorliegende Arbeit versteht sich als wissenschaftlicher Unterbau für die Entwicklung einer therapeutischen Mimiktraining-Plattform. Im ersten Unterkapitel dieser Einleitung wird ein Konzept für eine umfassende Trainingsplattform beschrieben und die Zielstellung dieser Arbeit darin verortet. Im Anschluss folgt eine detaillierte Übersicht über die konzeptionellen, technischen, methodischen und experimentellen Beiträge, sowie abschließend eine Übersicht über die Gliederung dieser schriftlichen Ausarbeitung.

1.1. Motivation und Zielstellung

Die Ursachen und Erscheinungsbilder von Fehlfunktionen der Gesichtsmuskulatur sind vielfältig und nicht an ein bestimmtes Lebensalter gebunden. Sie umfassen Lähmungen in Folge eines Schlaganfalls, aber auch Entwicklungsstörungen im Kindesalter, sogenannte Myofunktionelle Dysfunktionen [Kittel, 2008]. Letztere beschreiben Ungleichgewichte der Zungen- und Mundmuskulatur.

Die Auswirkungen für die Patienten sind ebenfalls vielfältiger Natur. Neben der Bedeutung der Mimik für die zwischenmenschliche Kommunikation, kann eine mangelnde Beherrschung der Gesichtsmuskulatur zu massiven Beeinträchtigungen des Alltags und gesundheitlichen Folgeschäden führen. Dazu zählen beispielsweise eine verwaschene Aussprache und Probleme bei der Nahrungsaufnahme. Des Weiteren vermindert ein unzureichender Lidschluss die natürliche Schutzfunktion des Auges vor Austrocknung und mechanischen Verletzungen. Betroffene Patienten müssen während des Schlafens einen Augenschutz tragen und kritische Freizeitbeschäftigungen, wie beispielsweise das Schwimmen, meiden [Brach und VanSwearingen, 1999].

Die Behandlung richtet sich nach der jeweiligen Ursache, die logopädische Therapie stellt jedoch im Allgemeinen einen wichtigen Teilaspekt des Rehabilitationsprozesses dar. Unter Anleitung eines Logopäden führt der Patient dabei Mimikübungen durch, die eine Reaktivierung der Gesichtsmuskulatur bewirken sollen. Die Mimikübungen werden im Folgenden auch als Fazialisübungen (*lat.* facies: Gesicht) bezeichnet. Ergänzende Trainingseinheiten ohne Aufsicht sind ebenfalls erforderlich. Ein Spiegel dient

(a) (b)

Abbildung 1.1.: **(a)** Mimikübungen mit Einsatz eines Spiegels zur Selbstkontrolle. **(b)** Szenario des Einsatzes einer Übungssoftware zum Mimiktraining. Der Patient erhält umgehend Feedback und kann seine Durchführung daran anpassen (© TU ILMENAU / Michael Reichel).

hierbei der Selbstkontrolle. Die Abbildung 1.1a verbildlicht ein solches Übungsszenario.

Die selbstständigen Trainingseinheiten sind jedoch nicht für jeden Patienten gleichermaßen geeignet und können bei falscher Durchführung kontraproduktiv wirken [Wolowski, 2005]. Zudem fehlen dem Logopäden ausreichende Informationen um die Durchführung und den Fortschritt der eigenständigen Übungseinheiten zu beurteilen. Der Einsatz einer kamerabasierten, automatisierten Trainingssoftware könnte den vom Logopäden unüberwachten, selbstständigen Teil des Übungsprozesses in vielerlei Hinsicht bereichern, wie in den folgenden Absätzen deutlich wird. Ein beispielhaftes Übungsszenario ist in der Abbildung 1.1b zu sehen. Der Patient sitzt dabei vor einer Kamera und führt Mimikübungen, gemäß eines vom Logopäden konzipierten Trainingsplans, aus. Die Aufnahmen des Patienten werden von der Trainingssoftware verarbeitet und automatisiert ausgewertet, um globales und regionenbezogenes Feedback für den Patienten zu generieren.

Die Entwicklung eines real einsetztbaren Systems ist jedoch ein komplexes Unterfangen, das zahlreiche Aufgabenstellungen umfasst. Zu diesen zählen unter anderem die Entwicklung der zu Grunde liegenden technischen Verfahren, die Laufzeitoptimierung, sowie die Usability. Aufgrund dieses Umfangs ist eine eingehende Betrachtung all dieser Aspekte im Rahmen der vorliegenden Arbeit nicht möglich.

Die **wesentliche Zielstellung dieser Arbeit** besteht in der Entwicklung, Implementierung und Evaluierung einer Methode zur Ableitung von globalem und regionenbezogenem Feedback aus den Tiefenaufnahmen von Patienten mit Mimikdysfunktionen.

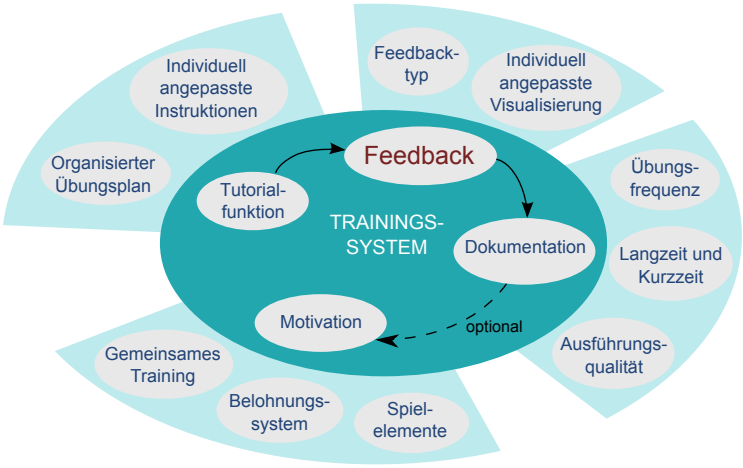


Abbildung 1.2.: Modell einer komplexen, umfassenden Trainingsplattform. Die vorliegende Arbeit konzentriert sich auf die Konzeption, Implementierung und Evaluation der Feedbackfunktionalität. Nähere Details zur den übrigen Funktionalitäten sind in [Dittmar et al., 2014] beschrieben.

Ein Ausblick für ein Konzept einer umfassenden Trainingsplattform, in welche sich ein entsprechendes Feedbacksystem integrieren ließe, ist in der Abbildung 1.2 gezeigt. Sie setzt sich aus vier Hauptfunktionalitäten zusammen, die im Folgenden in Kurzform beschrieben werden. Ausführlichere Informationen finden sich in [Dittmar et al., 2014].

- Die unüberwachten, selbstständigen Trainingseinheiten werden, gemäß eines vom Logopäden erstellten, individuellen Trainingsplans durchgeführt. Mit Hilfe einer integrierten *Tutorialfunktionalität*, die beispielsweise einen Zugriff auf anleitende Videos ermöglicht, könnte der Patient bei Bedarf die korrekte Ausführung der einzelnen Fazialisübungen nachvollziehen und rekapitulieren. Die Art und Weise der Informationsvermittlung ließe sich dabei an das Alter und die kognitiven Fähigkeiten des Patienten anpassen.

- Über die *Feedbackfunktionalität* der Trainingsplattform würde der Patient bereits während des Trainingsprozesses eine visuelle oder auditive Rückmeldung zu seiner Übungsausführung erhalten. Dies erfordert ein automatisiertes Verfahren, um die mit einer Kamera aufgenommenen Patientendaten zu verarbeiten und zu analysieren.

- Ebenfalls sinnvoll wäre die Integration einer *Dokumentationsfunktionalität*, die eine Nachvollziehbarkeit des Therapiefortschritts, auch über einen längeren Zeitraum hinweg, ermöglicht. Neben dem Übungserfolg ließen sich weitere Eckdaten, wie beispielsweise die Übungsfrequenz, dokumentieren.

Abbildung 1.3.: Thematische Grundpfeiler dieser Arbeit.

- Des Weiteren ließe sich die Trainingssoftware um *motivatorische Komponenten*, wie Belohnungssysteme oder spielerische Elemente, ergänzen. Entsprechende Untersuchungen zeigten auch bei Erwachsenen eine Motivationssteigerung durch den Einsatz virtueller Systeme, weshalb diese in der physischen und kognitiven Rehabilitation vermehrt Aufmerksamkeit erlangen (siehe z.B. [Taylor et al., 2011]).

Wie die im Abschnitt 2.1 dokumentierten Ergebnisse der Literaturrecherche verdeutlichen werden, steckt die Entwicklung von automatisierten Anwendungen zur Therapie von Mimikdysfunktionen noch in ihren Anfängen. Eine ausführliche Übersicht über den Beitrag der vorliegenden Arbeit zu dieser Thematik findet sich im folgenden Unterkapitel.

1.2. Beitrag der Arbeit

Die vorliegende Arbeit setzt sich im Wesentlichen aus drei thematischen Grundpfeilern zusammen, die in der Abbildung 1.3 gezeigt sind. Auf Basis dieser Grundpfeiler werden in den folgenden drei Absätzen die konzeptionellen, methodischen, technischen und experimentellen Beiträge dieser Arbeit vorgestellt.

Anwendungsszenario

Der Ausgangspunkt dieser Arbeit ist das Anwendungsszenario einer automatisierten, therapiebegleitenden Anwendung, welche im Rahmen der Rehabilitation von Mimikdysfunktionen einsetzbar ist. Um eine Einordnung des szenariobezogenen Beitrags dieser Arbeit in den Stand der Technik ermöglichen, findet sich zu Beginn dieser Ausarbeitung im Unterkapitel 2.1 eine ausführliche Recherche und Systematisierung

existierender wissenschaftlicher und praxisbezogener Ansätze. Die wissenschaftsbezogene Übersicht umfasst insgesamt 27 Publikationen aus den Jahren 1999 bis 2016. Ein deutliche Mehrheit davon befasst sich mit der Entwicklung von computergestützten bzw. automatisierten Verfahren zur Diagnose von Mimikdysfunktionen. Demgegenüber beschäftigt sich nur eine kleine Untermenge von fünf Veröffentlichungen mit der Entwicklung von automatisierten Therapiesystemen (siehe dazu Übersichtstabelle B.1 im Anhang). Zielstellung und Beitrag der vorliegenden Arbeit ordnen sich in die zweite Kategorie ein und weisen gegenüber den recherchierten Ansätzen folgende Vorteile auf:

- Das vorgestellte Verfahren erfordert lediglich einen Computer und einen RGB-D-Kamerasensor. Eine zusätzliche Anbringung von externen Schnittstellen oder Markern an Kopf oder Gesicht des Patienten ist nicht vonnöten (vgl. [Jayatilake et al., 2012]).

- Das entwickelte Verfahren stützt sich nicht auf die Analyse und Bewertung der Gesichtsasymmetrie. Dies bedeutet, dass auch achsenunsymmetrische Fazialisübungen in den Trainingsprozess einbezogen werden können (vgl. [Gebhard et al., 2000], [He et al., 2008], [Jayatilake et al., 2012]).

- Das entwickelte Verfahren ist universal gehalten. Während die zu Grunde liegenden Merkmalsextraktionsverfahren in [Y.-X. Wang et al., 2014] und [Tasneem et al., 2014] an konkrete Bewegungen angepasst sind (z.B. Beißbewegung, Augenöffnung), ist der in dieser Arbeit vorgestellte Ansatz flexibel auf weitere Fazialisübungen übertragbar. Eine Anpassung der Merkmalsextraktionsverfahren oder Datenauswertung ist dazu nicht erforderlich.

- Das vorgestellte Verfahren kann sowohl zur Erzeugung von globalem als auch regionenbezogenem Feedback eingesetzt werden.

Der Entwicklung und Umsetzung des Feedbackverfahrens liegen eine Vielzahl unterschiedlicher, teilweise aufeinander aufbauender, Aufgabenstellungen zu Grunde. Der folgende Absatz fasst diese methodischen und experimentellen Beiträge zusammen.

Methodischer und experimenteller Unterbau

Ein wesentlicher Beitrag dieser Arbeit besteht in der ausführlichen experimentellen Evaluation des entwickelten Feedbackverfahrens und der zu Grunde liegenden Teilkomponenten. Das Feedback selbst wird aus einem Merkmalsvektor abgeleitet, der

zuvor aus einer Tiefenbildaufnahme des Patienten extrahiert wurde. Da die Auswertung entsprechender therapiebezogener Literatur keine geeigneten Anknüpfungspunkte ergab, wird auf allgemeine Verfahren zur Gesichtsanalyse, beispielsweise aus der Emotionserkennung, zurückgegriffen.

Zu diesem Zweck wurde eine Literaturrecherche durchgeführt und insgesamt 28 Veröffentlichungen aus den Jahren 2006 bis 2014 systematisch ausgewertet. Da die Anzahl entsprechender Verfahren zur Gesichtsanalyse außerordentlich hoch ist, erhebt diese Übersicht keinen Anspruch auf Vollständigkeit und konzentriert sich auf merkmals- und tiefendatenbasierte, statische Extraktionsverfahren.

Auf Grundlage der Literaturrecherche wurden fünf Merkmalstypen (Extraktionsverfahren) ausgewählt, in Matlab nachimplementiert und evaluiert. Die Ergebnisse der Einzelevaluationen werden in Kapitel 4 beschrieben, die Ergebnisse der kombinierten Evaluation in Unterkapitel 5.3.

Zum Zweck der Feedbackableitung werden die extrahierten Merkmalsvektoren einem zuvor trainierten Random-Forest übergeben. Aus diesem lassen sich die paarweisen Ähnlichkeiten zwischen der Testobservation (Patientenobservation) und den Trainingsobservationen schätzen. Die paarweisen Ähnlichkeiten bilden die Grundlage für die Erzeugung des globalen und regionenbezogenen Feedbacks und werden in den Unterkapiteln 5.1 bis 5.3 vorgestellt und evaluiert.

Den bisher beschriebenen Analysen liegen manuell gesetzte Landmarken, beispielsweise zur Eingrenzung von Merkmalsextraktionsarealen, zu Grunde. Da ein manuelles Platzieren der Landmarken in einem automatisierten Szenario nicht sinnvoll ist, wird im Abschnitt 5.5.1 der Einfluss automatisiert lokalisierter Landmarken und Extraktionsareale auf die Ergebnisse der Random-Forest-Klassifikation untersucht.

Das Feedback in seiner endgültigen Form liegt in sechs diskreten Feedbackstufen vor. Die quantitative und qualitative Evaluation des lokalen und regionenbezogenen Feedbacks wird im Abschnitt 5.5.2 dokumentiert. Sie basiert sowohl auf Aufnahmen von gesunden Personen als auch auf Aufnahmen von Patienten mit Mimikdysfunktionen.

Prototyp

Um die Ergebnisse des entwickelten Verfahrens zu veranschaulichen, wurde ein Prototyp mit grafischer Benutzeroberfläche in Matlab umgesetzt (siehe Unterkap. 5.4). Neben einem Modellbild für die auszuführende Übung und der zu bewertenden Patientenaufnahme, umfasst die Benutzeroberfläche fünf Feedbackelemente, anhand derer sich verschiedene Aspekte des generierten globalen und lokalen Feedbacks nachvoll-

ziehen lassen. Der Schwerpunkt dieser Arbeit liegt jedoch auf der Entwicklung der Feedbackmethode, sowie der experimentellen Evaluation der zu Grunde liegenden Algorithmen. Der Prototyp dient somit allein der Ergebnisdemonstration und soll keine einsatzfähige Anwendung repräsentieren. Entsprechende weiterführende Aspekte wie laufzeitoptimierte Programmierung oder Benutzerfreundlichkeit sind daher nicht Gegenstand dieser Arbeit.

Fazit

In den vorhergehenden Absätzen wurden die wesentlichen Beiträge dieser Arbeit vorgestellt. Dabei wurde deutlich, dass sich der Entwicklungprozess des Feedbackverfahrens aus mehreren, teilweise aufeinander aufbauenden, Teilaufgaben zusammensetzt. Nähere Informationen zum Aufbau dieser Arbeit und der Beschreibung der Teilaufgaben sind im folgenden Unterkapitel zusammengefasst.

1.3. Gliederung

Die Gliederung dieser Arbeit orientiert sich weitestgehend chronologisch an der Struktur der technischen Gesamtarchitektur, welche dem entwickelten Feedbackverfahren zu Grunde liegt. Die Zuordnung der einzelnen Prozessschritte zu den Kapiteln dieser Arbeit wird aus der Abbildung 1.4 ersichtlich.

Der Beitrag von *Kapitel 2* ist im Wesentlichen konzeptioneller Natur. Einleitend findet sich eine Übersicht über den Stand der Technik bei der computergestützten Rehabilitation von Mimikdysfunktionen. Anschließend werden verschiedene Ursachen und Erscheinungsformen von Mimikdysfunktionen beschrieben. Das Unterkapitel 2.3 stellt die Datensätze vor, die im Rahmen der experimentellen Evaluationen Anwendung fanden. Das abschließende Unterkapitel enthält eine Übersicht über die technische Gesamtarchitektur und dient als Orientierungsrahmen für die folgenden Kapitel 3 bis 5, deren Fokus auf den einzelnen Komponenten der Architektur liegt.

Das *Kapitel 3* dokumentiert die Aufnahme und Vorverarbeitung der Bilddaten. Da die von der Kinect ausgegebenen RGB- und 2.5D-Bilder nicht deckungsgleich sind, ist die Kenntnis der Kameramatrix erforderlich, um korrespondierende Farb- und Tiefenpixel zu bestimmen. Sie ist desweiteren vonnöten, um die Tiefendaten aus dem 2.5D-Bild in eine 3D-Punktwolke zu überführen und wird im Rahmen der Kalibrierung bestimmt. Zwar bieten Funktionsbibliotheken, wie beispielsweise die *Point Cloud Library* oder die *Kinect SDK* ([Rusu und Cousins, 2011], [MICROSOFT-SDK]), Funktionen für die Ableitung von Punktwolken aus 2.5D-Bildern an, ihr Nachteil ist

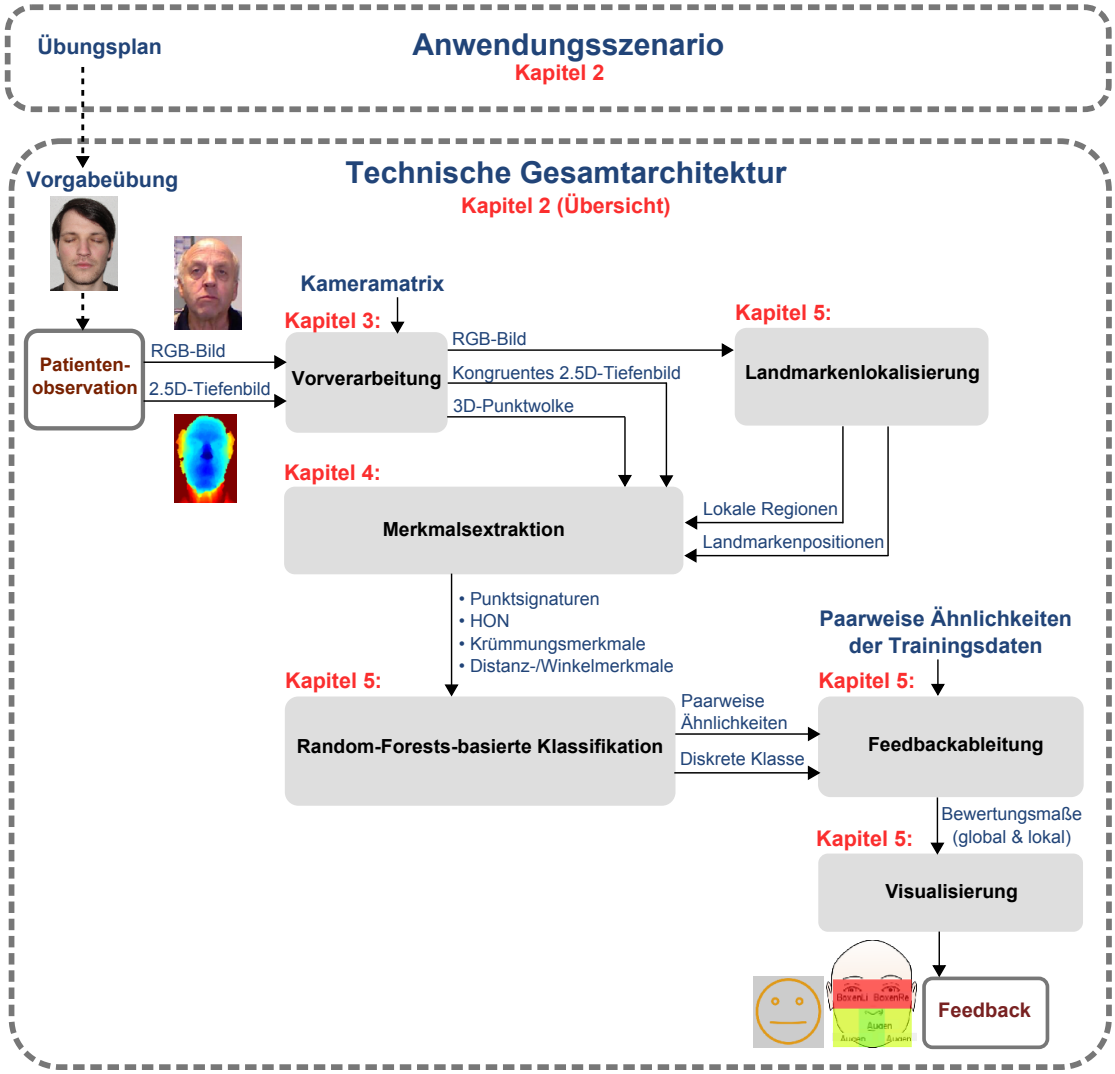


Abbildung 1.4.: Einbettung der Gliederung in das Anwendungsszenario und die technische Gesamtarchitektur.

jedoch, dass sie für die konkrete Aufgabenstellung überdimensioniert sind und weitere Abhängigkeiten, beispielsweise zu anderen Bibliotheken oder bestimmten Betriebssystemen, aufweisen. Aus diesem Grund wurde in dieser Arbeit eine entsprechende Funktion in Matlab nachimplementiert. Sie ist nicht an ein bestimmtes Betriebssystem gebunden und ohne weitere Abhängigkeiten flexibel einsetzbar. Die korrekte Vorverarbeitung der Tiefendaten ist entscheidend für ein funktionierendes Feedbacksystem. Die wesentliche Zielstellung von Kapitel 3 besteht darin, die ausgeführten Vorverarbeitungsschritte durchgehend nachvollziehbar zu machen. Mit Ausnahme des in Abschnitt 3.4 vorgestellten heuristischen Verfahrens, stützen sich diese auf existierende grundlegende Methoden.

Das *Kapitel 4* bildet, gemeinsam mit Kapitel 5, den Hauptteil dieser Arbeit. Ausge-

hend von einer systematischen Literaturauswertung im Unterkapitel 4.1 werden fünf geeignete Merkmalsextraktionsverfahren (Merkmalstypen) ausgewählt. Anschließend folgt ein Überblick über das allgemeine Testszenario, sowie, in den Unterkapiteln 4.3 bis 4.6, die gesonderte Beschreibung und experimentelle Evaluierung der einzelnen Extraktionsverfahren. In der Zusammenfassung findet sich eine Gegenüberstellung der Einzelergebnisse und ein Leitfaden zur Merkmalsauswahl.

Der Fokus des abschließenden *Kapitel 5* liegt auf der Methode zur Ableitung von globalem und lokalem Feedback aus den extrahierten Merkmalsdeskriptoren. Das Kapitel gliedert sich in zwei Teile. Der erste Teil beschäftigt sich mit der Entwicklung und Evaluierung des entwickelten Verfahrens und seiner zu Grunde liegenden Methoden. Der Schwerpunkt des zweiten Teils liegt auf der Auswertung des resultierenden diskreten Feedbacks und der Vorstellung des Prototypen zur Visualisierung des Feedbacks.

1.4. Publikationen

Vorarbeiten und Teile dieser Dissertation wurden bereits im Rahmen von Konferenzen und Buchbeiträgen publiziert. Darüber hinaus wurde auch an Veröffentlichungen mitgewirkt, welche nicht in direktem Bezug zu dieser Arbeit stehen. Im Folgenden findet sich eine entsprechende Auflistung.

Konferenzbeiträge mit Bezug zu dieser Arbeit

[Lanz et al., 2013a] Cornelia Lanz, Joachim Denzler und Horst-Michael Gross [2013a]. “Mimikdysfunktionen: Konzeption eines therapiebegleitenden Trainingssystems”. In: *Deutscher Ambient-Assisted-Living Kongress (AAL)*. Berlin, Germany, S. 186–195

> Der Fokus dieser Veröffentlichung liegt zum einen auf der Ausformulierung und Diskussion des Anwendungsszenarios. Zum anderen wird die Gesamtarchitektur des geplanten Systems vorgestellt und eine erste experimentelle Evaluation von gewählten Merkmalsextraktionsverfahren durchgeführt (vgl. dazu in dieser Arbeit Unterkap. 1.1 und 2.1).

[Lanz et al., 2013b] Cornelia Lanz, Joachim Denzler und Horst-Michael Gross [2013b]. “Robust landmark localization for facial therapy applications”. In: *European Conference on Technically Assisted Rehabilitation (TAR)*. Berlin, Germany

> In dieser Veröffentlichung wird ein krümmungs- und wissensbasiertes Verfahren zur automatisierten Lokalisierung von Landmarken im Gesicht vorgestellt und

evaluiert. Die Landmarkenlokalisierung bildet die Grundlage für die nachfolgende Extraktion der Merkmale. Ein Überblick über die untersuchten Ansätze und das letztendlich gewählte Verfahren findet sich in dieser Arbeit in Abschnitt 5.5.1.

[Lanz et al., 2013c] Cornelia Lanz, Birant Sibel Olgay, Joachim Denzler und Horst-Michael Gross [2013c]. "Automated classification of therapeutic face exercises using the kinect". In: *International Conference on Computer Vision Theory and Applications (VISAPP)*. Barcelona, Spain, S. 556–565 - **Best Student Paper Award**

Der Schwerpunkt dieser Veröffentlichung liegt auf den Merkmalsextraktionsverfahren, bestehend aus Punktsignaturen, Linienprofilen und Krümmungsmerkmalsdeskriptoren. Die Extraktionsverfahren werden vorgestellt und detailliert evaluiert (vgl. dazu in dieser Arbeit Kap. 4 und Unterkap. 5.3). Zudem wird ihre Robustheit gegenüber automatisiert lokalisierten Landmarkenpositionen untersucht (vgl. Abschn. 5.5.1). Die Landmarkenlokalisierung basiert dabei auf sogenannten Active-Appearance-Models.

[Dittmar et al., 2017] Cornelia Dittmar, Joachim Denzler und Horst-Michael Gross [2017]. "A feedback estimation approach for therapeutic facial training". In: *12th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. Accepted. Washington, D.C., USA

Diese Veröffentlichung beschreibt das in dieser Arbeit entwickelte Feedbackverfahren. Der Schwerpunkt liegt hierbei auf den methodischen Aspekten und der qualitativen Evaluation (vgl. dazu in dieser Arbeit Kap. 5) .

Buchbeiträge mit Bezug zu dieser Arbeit

[Lanz et al., 2014] Cornelia Lanz, Birant Sibel Olgay, Joachim Denzler und Horst-Michael Gross [2014]. "Facial Landmark Localization and Feature Extraction for Therapeutic Face Exercise Classification". In: *Computer Vision, Imaging and Computer Graphics – Theory and Applications*. Communications in Computer and Information Science. Springer, S. 179–194. DOI: 10.1007/978-3-662-44911-0

Dieser Beitrag ist Teil einer Buchveröffentlichung von ausgewählten und überarbeiteten Publikationen der *VISAPP 2013*. Im Vergleich zur Konferenzversion wird ein weiteres Landmarkenlokalisierungsverfahren vorgestellt und mit der AAM-basierten Lokalisierung verglichen. Es handelt sich dabei um einen kombinierten Ansatz, welcher den auf der *TAR 2013* vorgestellten Lokalisierungsansatz und sogenannte *tree-structured parts models* vereinigt.

[Dittmar et al., 2014] Cornelia Dittmar, Joachim Denzler und Horst-Michael Gross [2014]. "Facial movement dysfunctions: Conceptual design of a therapy-accompanying training system". In: *Ambient Assisted Living*. Advanced Technologies and Societal Change. Springer Berlin Heidelberg, S. 123–141. DOI: 10.1007/978-3-642-37988-8_9

Zusammenfassung und Veröffentlichung der Publikationen des *AAL-Kongresses 2013* in Form eines Buches.

Weitere Veröffentlichungen

[Dunker et al., 2008] Peter Dunker, Stefanie Nowak, André Begau und Cornelia Lanz [2008]. "Mood classification for photos and music: A generic multi-modal classification framework and evaluation approach". In: *ACM International Conference on Multimedia Retrieval (MIR)*. Vancouver, Canada, S. 97–104

[Lanz et al., 2010b] Cornelia Lanz, Stefanie Nowak und Uwe Kuehhirt [2010b]. "Determination of categories for tagging and automated classification of film scenes". In: *European Conference on Interactive TV and Video (EuroITV)*. Tampere, Finland, S. 297–300

[Lanz et al., 2010a] Cornelia Lanz, Hanna Lukashevich und Stefanie Nowak [2010a]. "Automated classification of film scenes based on film grammar". In: *Workshop Audiovisuelle Medien (WAM)*. Chemnitz, Germany, S. 143–155

2. Anwendungsszenario und Gesamtarchitektur

Die Zielstellung dieses Kapitels ist konzeptioneller und einleitender Natur. Im ersten Unterkapitel werden existierende wissenschaftliche und praxisbezogene Ansätze zur computergestützten Rehabilitation von Mimikdysfunktionen vorgestellt und der konzeptionelle Beitrag dieser Arbeit herausgearbeitet. Anschließend folgt ein Überblick über die verschiedenen Ursachen und Erscheinungsformen von Mimikdysfunktionen. Im Unterkapitel 2.3 findet sich eine Beschreibung der Datenbasis, die der Entwicklung und experimentellen Evaluation des Feedbackverfahrens zu Grunde liegt. Abschließend wird eine Übersicht über die technische Gesamtarchitektur des entwickelten Verfahrens gegeben. Sie dient als Orientierungsrahmen für die folgenden Kapitel, die sich mit einzelnen Teilaspekten der Gesamtarchitektur beschäftigen.

2.1. Computergestützte Therapieansätze

Computergestützte Lösungen gewinnen in der Rehabilitation kontinuierlich an Bedeutung. Der Markteintritt der *Wii-Konsole* von Nintendo und des *Kinect-Sensors* von Microsoft stellen wichtige Startpunkte dieser Entwicklung dar ([NINTENDO-Wii], [MICROSOFT-XBox]). Beide Systeme wurden ursprünglich für den Consumer-Spielemarkt entwickelt und finden nun in der physischen Rehabilitation, unter anderem bei der Nachbehandlung von Schlaganfällen, Anwendung. Ein Beispiel hierfür ist die Kinect- und webbasierte Softwareplattform *Jintronix* [JINTRONIX]. Sie enthält verschiedene Bewegungsspiele und übermittelt die in den Trainingsspielen gewonnenen Bewertungsparameter wie Schnelligkeit, Präzision und Bewegungsumfang an den Therapeuten, welcher auf diese Weise den Fortschritt seiner Patienten kontrollieren kann. Verglichen mit der physischen Rehabilitation sind zum gegenwärtigen Zeitpunkt computergestützte Verfahren in der Behandlung von Mimikdysfunktionen weniger etabliert. Ein Überblick über den Forschungsstand der computergestützten Fazialisparese-Rehabilitation findet sich in Abschnitt 2.1.1. Beispiele für kommerzielle Anwendungen sind im Abschnitt 2.1.2 dokumentiert.

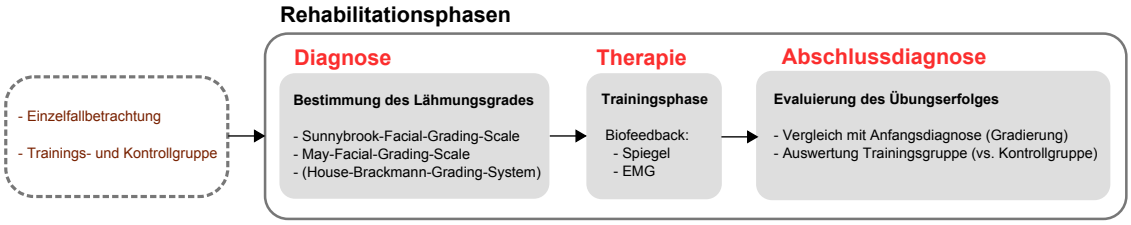


Abbildung 2.1.: Schematischer Rehabilitationsablauf gemäß [Brach und VanSwearingen, 1999] und [Cronin und Steenerson, 2003].

2.1.1. Stand der Forschung

Die folgende Literaturübersicht stützt sich auf die systematische Auswertung von insgesamt 27 Publikationen aus den Jahren 1999 bis 2016. Ihr Fokus liegt auf dem Anwendungsszenario der Fazialisparese-Rehabilitation, da sich die recherchierten automatisierten Ansätze zur Rehabilitation der myofunktionellen Dysfunktion im Wesentlichen auf die Analyse von Sprachsignalen beziehen. Das Ziel der Literaturübersicht ist ein allgemeiner und systematischer Überblick über die Ergebnisse der Recherche. Tiefergehende Details zu den einzelnen Veröffentlichungen finden sich im Anhang in der Tabelle B.1.

Den zahlreichen Ursachen der Fazialisparese entsprechend, existieren verschiedene Rehabilitationsansätze. Neben der operativen oder medikamentösen Behandlung, wurde und wird intensiv der Nutzen von Mimiktraining untersucht. Wissenschaftliche Studien zum nicht-computergestützten therapeutischen Mimiktraining finden sich unter anderem in [Brach und VanSwearingen, 1999] und [Cronin und Steenerson, 2003]. Der Ablauf beider Studien ist in der Abbildung 2.1 schematisch visualisiert. Aus ihm lassen sich zwei zentrale Bestandteile der Rehabilitation identifzieren: die *Diagnose* und die *Therapie*.

Zu Beginn der Rehabilitation erfolgt im Rahmen der Diagnose die Beurteilung der mimischen Funktionen des Patienten. Diese umfassen beispielsweise die Fähigkeit zum Lidschluss, den generellen Funktionsumfang, sowie die Ruhe- und Bewegungsasymmetrie. Um eine Objektivierung dieses Prozesses zu erreichen, wurden Gradierungssysteme entwickelt. Zu den Bekanntesten zählen unter anderem das House-Brackmann-Grading-System (HBGS) und das Sunnybrook-Facial-Grading-System (SFGS) ([House und Brackmann, 1985], [Ross et al., 1996]). Eine Analyse und Gegenüberstellung verschiedener Gradierungssysteme findet sich in [Zhai et al., 2008]. An die Diagnose schließt sich eine mehrere Wochen oder Monate andauernde Therapiephase an, welche die regelmäßige Ausübung von Mimikübungen beinhaltet. Dabei ist auf eine korrekte Ausführung zu achten, um beispielsweise die Entwicklung von Synkinesien zu ver-

meiden ([Kleiss et al., 2013], [Wolowski, 2005], [Nakamura et al., 2003]). Unter Synkinesien versteht man fehlgeleitete Neuverknüpfungen von Nervenenden, die zu unwillkürlichen Mitbewegungen im Zuge von beabsichtigten Mimikbewegungen führen können. Dies kann sich beispielsweise in Form eines nicht willkürlich gesteuerten Lidschlusses während einer Bewegung der Lippen äußern.

Insbesondere im Frühstadium der Fazialisparese, wenn die Aktivierung der Muskeln noch nicht in einer sichtbaren Bewegung resultiert, kann der unterstützende Einsatz der Elektromyographie sinnvoll sein, um dem Patienten Feedback zur Verfügung zu stellen [Brach und VanSwearingen, 1999]. Bei sichtbaren Mimikbewegungen ist eine Beaufsichtigung durch einen Logopäden oder eine Selbstkontrolle mittels Spiegel möglich. Auf Basis einer erneuten Gradierung lässt sich, sowohl während als auch nach Abschluss der Therapie, der (bisherige) Behandlungserfolg quantifizieren.

Gradierungssysteme, wie beispielsweise das HBGS und das SFGS, leisten bereits einen Beitrag zur Objektivierung der Diagnose. Da sie dennoch in gewisser Weise auf der subjektiven Einschätzung der behandelnden Person basieren, wird in den ausgewerteten Veröffentlichungen eine weitere Objektivierung des Diagnoseprozesses angestrebt. Die Fazialisparese äußert sich in der Regel sowohl in einer Ruhe- als auch einer Bewegungsasymmetrie zwischen beiden Gesichtshälften, weshalb die Mehrheit der Verfahren darauf ausgerichtet ist, die Ausprägung der statischen und dynamischen Asymmetrien zu quantifizieren. Detaillierte Beschreibungen zu den Inhalten der einzelnen Veröffentlichungen finden sich im Anhang in der Tabelle B.1.

Wie die Abbildung 2.2 zeigt, basieren insgesamt 15 von 22 Verfahren auf der Analyse von statischen Einzelbildern oder Punktwolken. Bei sieben Verfahren liegen Videodaten zugrunde, wobei dennoch häufig statische und dynamische Merkmale kombiniert werden. Im Rahmen der statische Merkmalsextraktion werden Landmarkenabstände ([Nakamura et al., 2003], [Dong et al., 2008], [Hadlock und Urban, 2012], [Kleiss et al., 2013], [Kim et al., 2015]), Intensitätswertdifferenzen ([S. Wang und Qi, 2005], [T. Wang et al., 2015], [Gebhard et al., 2001]) und Local-Binary-Patterns ([He et al., 2008], [T. Wang et al., 2015]) der gesunden und der betroffenen Gesichtshälfte gegenübergestellt. Andere Ansätze kombinieren die von Active-Appearance-Modellen abgeleiteten Parameter und Action-Units zur Quantifizierung des Fazialisparesegrades ([Haase et al., 2015], [Modersohn und Denzler, 2016]). Die Analyse der Videos umfasst desweiteren die Verfolgung der Landmarken über mehrere Frames [Delannoy und Ward, 2010], sowie die Extraktion des Optical-Flow ([Gebhard et al., 2001], [He et al., 2007]). Eine kombinierte raum- und zeitbezogene Extraktion von Local-Binary-Patterns findet sich in [He et al., 2009].

Durch die zunehmende Verfügbarkeit von kostengünstigen 3D-Sensoren, wächst

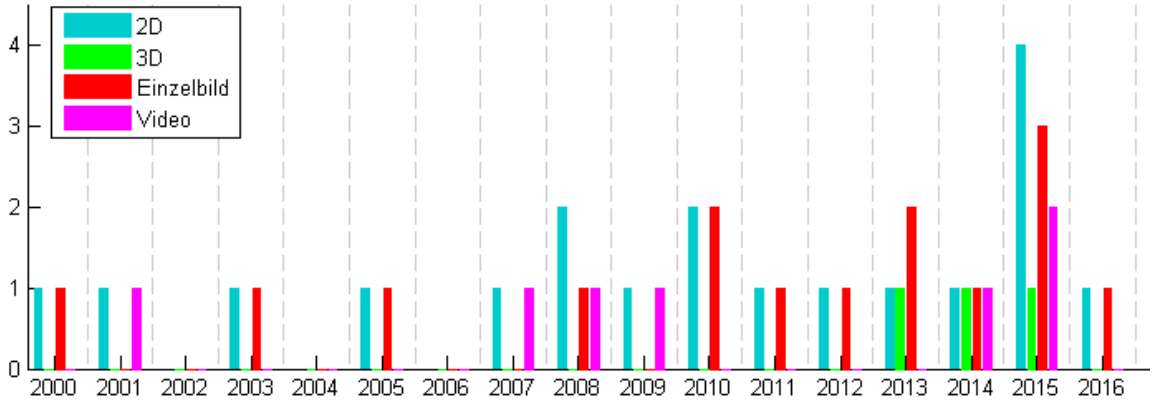


Abbildung 2.2.: Einordnung von 22 therapie- und diagnosebezogenen Veröffentlichungen entsprechend ihres Erscheinungsjahres und ihrer Datenbasis (2D vs. 3D, Einzelbild vs. Video). Die Visualisierung ist auf die teil- und vollautomatisierten Verfahren beschränkt. Das in [Jayatilake et al., 2012] beschriebene, teilautomatisierte, Verfahren ist nicht Bestandteil des Diagramms, da es auf Daten basiert, die mittels Elektromyographie erfasst wurden. Details zu allen Publikationen finden sich im Anhang in der Tabelle B.1.

auch die Anzahl entsprechender Verfahren. Das Ziel des in [Gaber et al., 2015] beschriebenen Ansatzes ist die Analyse der Ruhesymmetrie des Gesichts unter Einsatz eines mit der Kinect aufgenommenen Datensatzes von zehn gesunden Personen. Zu diesem Zweck wird für die Augenbrauen, die Augen und den Mund jeweils ein Symmetrieindex geschätzt. In diese Schätzung einbezogen werden im Wesentlichen die Schiefe des Mundes, die mittels einer eingepassten Ellipse angenäherten Flächeninhalte der Augenöffnungen, sowie die dreidimensionalen euklidischen Abstände zwischen definierten Landmarken. In [L. Y. Lin, 2013] wird ein binäres Fazialisparese-Diagnosesystem experimentell untersucht. Die Merkmalsextraktion umfasst, neben der Schätzung der Gauß'schen Krümmung und des Shape-Indexes, die Berechnung der geometrischen Distanz zwischen der linken und rechten Gesichtshälfte, welche zuvor anhand des Iterative-Closest-Point-Algorithmus miteinander registriert wurden.

Die Mehrheit der ausgewerteten Ansätze konzentriert sich auf das Szenario der Diagnose, wie anhand des Säulendiagramms in der Abbildung 2.3 ersichtlich wird. Mit der zunehmenden Robustheit und Genauigkeit der Landmarkenlokalisierungsverfahren, wächst jedoch auch die Anzahl der therapiebezogenen Ansätze. Während die Durchführung der Diagnose in der Regel auch offline realisierbar ist, setzt ein Therapieszenario eine weitestgehende Echtzeitperformance voraus. Zeitintensives und manuelles Positionieren von Landmarken würde den Übungsablauf dementsprechend wesentlich einschränken. In [Gebhard et al., 2000] werden aus dem Umfeld der äu-

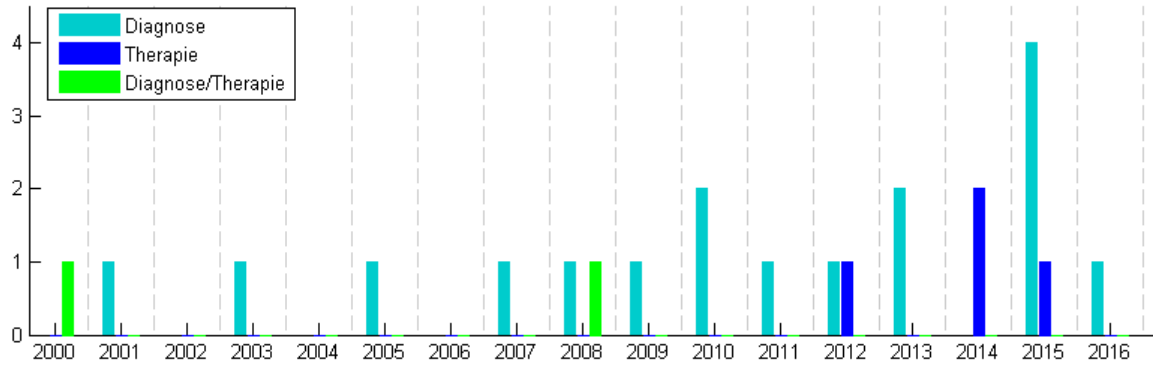


Abbildung 2.3.: Einordnung von 23 teil- und vollautomatisierten Verfahren entsprechend ihres Themenschwerpunktes und dem Jahr ihrer Veröffentlichung.

ßeren Augen- und Mundwinkel sogenannte Signaturen extrahiert. Sie stellen im Wesentlichen die Filterantworten von orientierbaren zweidimensionalen Gaußfiltern dar. Die Korrelation zwischen den Signaturen von gegenüberliegenden Landmarken dient als Maß für die Schätzung der Asymmetrie zwischen der linken und rechten Gesichtshälfte. Die Autoren empfehlen den Einsatz des geschätzten Asymmetriewertes sowohl für die Diagnose, als auch für den Einsatz in einem Therapieszenario.

Der Nachteil der (A-)Symmetrieanalyse besteht jedoch darin, dass sie die in Frage kommenden Krankheitsbilder für die Therapieanwendung einschränkt. Myofunktionelle Dysfunktionen beispielsweise zeichnen sich, anders als unilaterale Fazialisparesen, nicht vorwiegend durch Asymmetrie aus. Zudem schränkt die symmetriebezogene Auswertung die möglichen therapiebezogenen Mimikübungen auf achsensymmetrische Übungen ein. Nähere Details zu den verschiedenen Mimikdysfunktionen und den in dieser Arbeit gewählten therapeutischen Mimikübungen finden sich in den Abschnitten 2.2 und 2.3.1.

In [Tasneem et al., 2014] und [Y.-X. Wang et al., 2014] wird jeweils ein Prototyp für ein spielbasiertes Therapiekonzept vorgestellt. In Hinblick auf den erfolgreichen Einsatz von Spielkonsolen in der allgemeinen physischen Rehabilitation erscheint die Einbindung von spielerischen Elemente grundsätzlich vorteilhaft und sinnvoll. Die genannten Systeme funktionieren derzeit allerdings nur für einzelne spezifische Übungen (Lidschluss [Tasneem et al., 2014]; Beißbewegung, Zunge zeigen [Y.-X. Wang et al., 2014]). Die zugrunde liegenden Verfahren sind teilweise wissensbasiert und daher weniger flexibel erweiterbar als eine auf (überwachten) Lernverfahren basierende Methode.

2.1.2. Existierende praktische Anwendungen

Im kommerziellen Markt finden sich bereits einfache, nicht-automatisierte Softwareapplikationen zur Unterstützung des Mimiktrainings ([COMUZU], [NINTENDO-FACE], [Schreiber, 2015]). Ihr Beitrag liegt jedoch vielmehr in der Motivation und Vermittlung der Übungen. Eine Bewertung des Trainingserfolgs, welche über eine Selbstbeobachtung des Übenden hinausgeht, findet nicht statt.

2.1.3. Zusammenfassung und Bogen zu dieser Arbeit

Die im Rahmen der Literaturrecherche beschriebenen medizinischen Fallstudien und computergestützten Diagnoseverfahren stellen nur eine Teilmenge der existierenden, zahlreichen Ansätze dar. Demgegenüber beschränken sich die gefundenen vollautomatisierten Therapieansätze, trotz eingehender Suche, auf wenige Publikationen. Einige davon quantifzieren die Asymmetrie zwischen den Gesichtshälften und verwenden diese als Bewertungsmaß ([Gebhard et al., 2000], [He et al., 2008], [Tasneem et al., 2014], Details siehe Tab. B.1). Auf diese Weise ist ihr Einsatz auf die Therapie von unilateralen Dysfunktionen fokussiert. Für andere Krankheitsbilder, wie die myofunktionelle Dysfunktion, sind sie nur bedingt geeignet (siehe Abschn. 2.2). Einige der in dieser Arbeit gewählten therapeutischen Fazialisübungen sind nicht achsensymmetrisch und erfordern daher eine von Symmetrieannahmen unabhängige Vorgehensweise. Hierzu zählt das in [Y.-X. Wang et al., 2014] beschriebene interaktive Spiel. Ein weiterer Vorteil des Konzepts besteht darin, dass der Patient während des Spiels Punkte sammeln kann und dadurch ein Bewertungsmaß für seinen Übungserfolg erhält. Nachteilig ist jedoch, dass das Verfahren auf zwei Übungen (Beißbewegung, Zunge zeigen) beschränkt ist. Ziel und Beitrag dieser Arbeit bestehen in der Entwicklung eines therapietauglichen Feedbackverfahrens, welches ein quantifiziertes Bewertungsmaß ausgibt, unabhängig von festen Symmetrieannahmen ist und sich flexibel auf verschiedenste Mimikübungen übertragen lässt.

2.2. Mimikdysfunktionen und ihre Therapie

Die Gründe für Mimikdysfunktionen sind vielfältig. Sie reichen von einer mangelhaft und einseitig trainierten Muskulatur, auch als myofunktionelle Dysfunktion bezeichnet, hin zu vollständigen Lähmungen. In Abhängigkeit von der Ursache unterscheidet man bei einer Gesichtslähmung (*med.* Fazialisparese) zwischen einer peripheren, zentralen oder idiopathischen Parese [Berghaus et al., 1996]. Die *periphere Lähmung* ist auf eine Läsion des Gesichtsnervs (*med.* Fazialisnerv) zurückzuführen. Die Funktion

dieses Nervs besteht in der motorischen Innervation der Gesichtsmuskulatur, weshalb seine Endäste an die mimische Muskulatur anknüpfen [Probst et al., 2008]. Unter Innervation versteht man die "Versorgung von Geweben und Organen mit Nerven", sowie die "Leitung der Reize durch die Nerven (...)" [DUDEN]. Der Verlauf des Fazialisnervs ist in der Abbildung 2.4 zu sehen. Die möglichen Auslöser einer peripheren Gesichtslähmung sind ([Berghaus et al., 1996], [Probst et al., 2008], [Schwenkreis, 2012]):

- Tumore, die durch Quetschungen die Reizleitung des Fazialisnervs beeinträchtigen.
- Schädigungen oder Durchtrennungen des Fazialisnervs, die aus äußerer Gewalt hervorgehen. Diese Verletzungen werden auch als Traumen bezeichnet und können durch Unfälle oder Operationen im Bereich des Fazialisnervs enstehen.
- durch Viren oder Bakterien ausgelöste Entzündungen (z.B. Herpes zoster oticus, Borreliose). Erkrankungen des Ohres oder der Ohrspeicheldrüse können aufgrund der räumlichen Nähe ebenfalls auf den Fazialisnerv übergreifen.
- stoffwechselbedingte Ursachen. So tritt eine periphere Gesichtslähmung gehäuft bei Diabetes oder während der Schwangerschaft auf. Die genauen metabolischen Abläufe, die zu den Lähmungserscheinungen führen, sind jedoch noch nicht bekannt.

In 75% der Fälle lässt sich der peripheren Fazialisparese jedoch keine genaue Ursache zuordnen [Schwenkreis, 2012]. Man spricht dann von der sogenannten *idiopathischen Gesichtslähmung* als Ausschlussdiagnose [Probst et al., 2008]. Als mögliche Ursachen werden gefäßverändernde Ursachen wie Diabetes, Hypertonus oder Virusinfektionen diskutiert [Berghaus et al., 1996].

Bei der *zentralen Fazialisparese* ist der Gesichtsnerv nicht betroffen. Sie ist stattdessen auf eine Schädigung im Hirn zurückzuführen, beispielsweise in Folge eines Schlaganfalls oder eines Tumors ([Berghaus et al., 1996], [Schwenkreis, 2012]).

Neben einer Lähmung kann auch eine unzureichend oder falsch ausgebildete Gesichtsmuskulatur zu einer Beeinträchtigung der Mimik führen. Man spricht in diesen Fällen von einer myofunktionellen Störung. Der Name leitet sich vom griechischen Wort *mys* (*dt.* Muskel) ab und meint übersetzt *die Funktion des Muskels betreffend*. Mögliche Ursachen für diese Störung sind unter anderem ein falsches Schluckmuster oder Fehlbildungen wie Lippen-Kiefer-Gaumenspalten [Kittel, 2008].

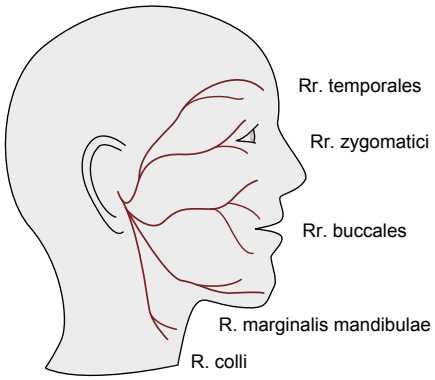


Abbildung 2.4.: Der Fazialisnerv teilt sich in fünf Endäste auf, die unterschiedliche Gesichtsmuskeln motorisch innervieren. Die lateinische Namensgebung der Endäste basiert auf den Gesichtsbereichen in die sie münden (Schläfe, Jochbein, Wange, Unterkiefer, Hals). Weitere Informationen zum Fazialisnerv finden sich im Anhang A.1.

Die Rehabilitation der Mimikdysfunktion richtet sich nach der jeweiligen Ursache. Bei idiopathischen Fazialisparesen erfolgt die Behandlung durch die Gabe von entzündungshemmenden Medikamenten und einem logopädischen Training der mimischen Muskulatur. Bei symptomatischen Fazialisparesen wird zuerst mit Blick auf die Ursache, z.B. einer Virusinfektion, behandelt. Logopädische und physiotherapeutische Maßnahmen können ebenfalls unterstützend eingesetzt werden [Schwenkreis, 2012]. Im Falle der myofunktionellen Dysfunktion sind diese Maßnahmen nicht nur unterstützender Natur sondern zentrale Behandlungsmethode [Kittel, 2008].

Bei der logopädischen Therapie wird eine Reaktivierung der Gesichtsmuskulatur durch Mimikübungen angestrebt. Die Durchführung erfolgt im Allgemeinen unter Anleitung einer Fachperson, ergänzt durch eigenständige und unbeaufsichtige Übungseinheiten vor einem Spiegel. Details dazu und zur Einordnung dieser Arbeit wurden bereits in den Unterkapiteln 1.1 und 1.2 beschrieben. Die Abbildung 2.5 zeigt die Ausführung zweier Mimikübungen durch einen Patienten mit einer rechtsseitigen Fazialisparese [Dittmar et al., 2014]. Beide Aufnahmen sind Teil des finalen Testdatensatzes, der im folgenden Unterkapitel vorgestellt wird.

2.3. Datenbasis dieser Arbeit

Das Ziel dieser Arbeit besteht in der Entwicklung, Implementierung und Evaluierung eines Verfahrens für das interaktive Mimiktraining. Die diesem Prozess zugrunde liegende Datenbasis ist Gegenstand der beiden folgenden Abschnitte. Im ersten Ab-

Abbildung 2.5.: Patient mit einer rechtsseitigen Fazialisparese. Linke Abbildung: Das Aufblasen der rechten Wange gelingt, da die Wölbung der Wange einen passiven Prozess darstellt. Sie ist die Folge eines erhöhten Luftdrucks im Mund und der Kontraktion der Wangenmuskulatur in der linken Gesichtshälfte. Rechte Abbildung: Die linke Wange kann nicht aufgeblasen werden, da keine Kontraktion in der rechten Wangenmuskulatur aufgebaut werden kann.

schnitt werden zwölf Fazialisübungen festgelegt, die für den Einsatz innerhalb eines therapeutischen Trainingssystems geeignet sind. Der zweite Abschnitt beschreibt die Zusammenstellung der Ground-Truth-Daten.

2.3.1. Wahl der therapeutischen Übungen

Im Zusammenarbeit mit der *Fachklinik Bad Liebenstein*[1] und der *Logopädischen Praxis Irina Stangenberger*[2] wurden zu Beginn dieser Arbeit insgesamt 19 klassische Fazialisübungen (*lat.* facies: Gesicht) ausgewählt, die für die Reaktivierung der Augen-, Zungen-, Mund- und Wangenmuskulatur geeignet sind [Okreu und Beckers, 2006]. Eine entsprechende Übersicht findet sich im Anhang in der Abbildung B.1.

Die Experimente in den Kapiteln 4 und 5 stützen sich der Übersichtlichkeit halber auf einen reduzierten Satz von zwölf Übungsklassen. Demgegenüber basieren die, in der Literaturübersicht im Unterkapitel 4.1 aufgeführten, vergleichbaren Referenzverfahren zur Mimikanalyse auf maximal sechs Basisemotionen und einem Neutralgesicht. Der gewählte Übungssatz beinhaltet alle elf Mund- und Wangenmuskulaturübungen, sowie eine Augenübung. Zehn Übungen sind in den Abbildungen 2.6a bis 2.6j gezeigt, die fehlenden zwei Fazialisübungen ergeben sich aus der achsengespiegelten Ausführung von *WangeLi* und *BoxenLi*.

Alle Übungen sind primär statischer Natur und sollten für einige Sekunden im ma-

[1] Fachklinik Bad Liebenstein, Abteilung *Weiterführende Neurorehabilitation*, Chefarzt Prof. Dr. med. Gustav Pfeiffer (Stand 2014), Bad Liebenstein. Internetseite: www.fachklinik-bad-liebenstein.de, letzter Zugriff: 16.01.2016.

[2] Logopädischen Praxis Irina Stangenberger, Arnstadt. Internetseite: www.logopaedie-stangenberger.de, letzter Zugriff: 16.01.2016.

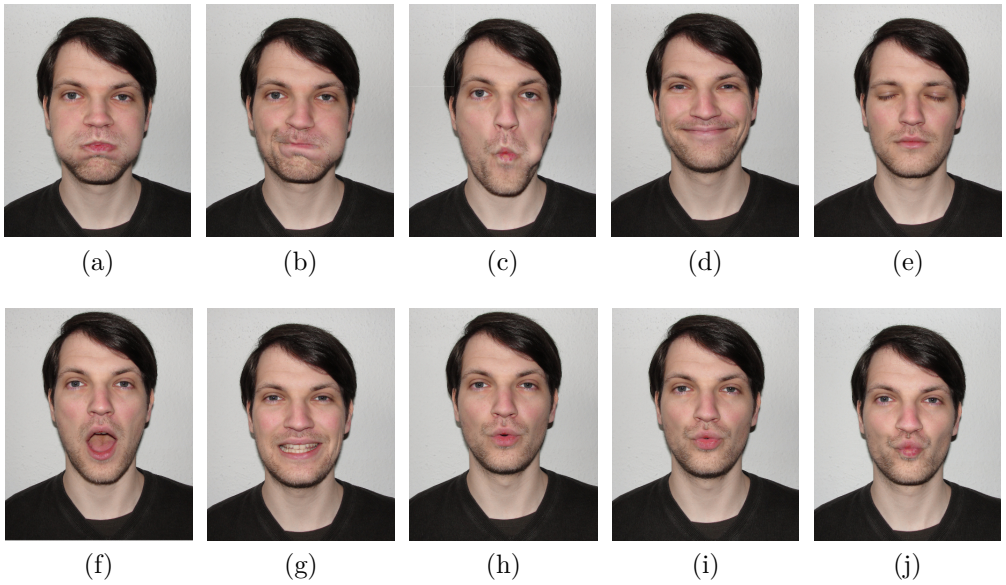


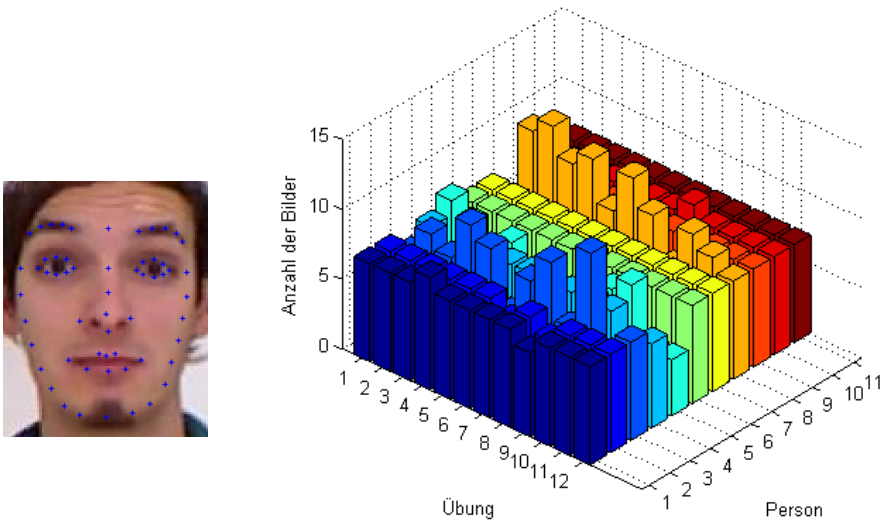
Abbildung 2.6.: Zehn der zwölf ausgewählten therapeutischen Mimikübungen. **(a)**–**(c)** Die Wangenübungen *Wangen*, *WangeLi*, *BoxenLi*. Die letzten beiden Übungen werden auch in achsengespiegelter Form für die rechte Wange ausgeführt (*WangeRe*, *BoxenRe*). Die Bezeichnung der Seite ist aus Patientensicht definiert. **(d)** Die Mundübung *Breit*. **(e)** Die Übung *Augen* entspricht in der unteren Gesichtshälfte einem entspannten Neutralgesicht. **(f)**–**(i)** Vier Vokalübungen *AForm*, *IForm*, *OForm* und *UForm*. **(j)** Die Übung *Kuss* weist eine hohe Ähnlichkeit zu den Vokalübungen *OForm* und *UForm* auf.

ximalen Ausführungszustand gehalten werden. Die Geschwindigkeit des Übergangs vom Neutralgesicht zum Übungsklimax ist dabei nicht relevant. Bei Bedarf ist jedoch auch ein dynamisches Training realisierbar, beispielsweise durch die alternierende Ausführung von zwei oder mehr Übungen. Eine mögliche Variante bildet der periodische Wechsel zwischen den Übungen *UForm* und *IForm*.

2.3.2. Ground-Truth-Daten

Die *Ground-Truth* ist ein wichtiger Bestandteil der Entwicklung und Evaluierung des Trainingssystems. Sie repräsentiert eine, in der Regel mit Metainformationen angereicherte, Sammlung von Referenzdaten. Die Metainformationen werden auch als Annotationen bezeichnet. Da über die Literaturrecherche keine geeignete Ground-Truth ausfindig gemacht werden konnte, wurden zwei Datensätze erstellt, die im Folgenden detaillierter beschrieben werden. Die Datenaufnahme und Annotation wurde durch die Mithilfe von Birant Sibel Olgay im Rahmen ihrer Masterarbeit [Olgay, 2012][3]

[3]Diese Masterarbeit wurde von der Autorin im Rahmen dieser Dissertation betreut.



(a) (b)

Abbildung 2.7.: **(a)** Positionen der 58 manuell annotierten Landmarken. **(b)** Bivariates Histogramm über die Aufteilung der 931 2.5D-Bilder auf elf Personen und 12 Übungen.

unterstützt.

Der erste Datensatz umfasst Aufnahmen von 19 verschiedenen Fazialisübungen, die jeweils von drei weiblichen und acht männlichen Personen im Alter von 25 bis 28 Jahren ausgeführt werden. Unter Einsatz der *Kinect for Xbox 360* von *Microsoft* wurden pro Person und Übung im Schnitt sieben Bilder aufgenommen ([MICROSOFT-XBox]). Die gesamte Ground-Truth setzt sich aus jeweils 1485 RGB- und 2.5-D-Tiefenbildern zusammen. Alle RGB-Bilder des Datensatzes wurden, unter Einsatz der Annotationssoftware *AAM Label Tool* des Fachgebiets Neuroinformatik und Kognitive Robotik der TU Ilmenau, manuell annotiert. Die Annotation erfolgte durch zwei Personen und umfasst die Identität der dargestellten Person, die ausgeführte Übung, sowie die Position von 58 definierten Landmarken innerhalb des Gesichts. Diese orientieren sich an den Landmarken der Active-Appearance-Models und sind in der Abbildung 2.7a gezeigt [Cootes et al., 2001].

Da von den insgesamt 19 Fazialisübungen eine Untermenge von zwölf für die Experimente verwendet wird, reduziert sich der Datensatz von 1485 auf jeweils 931 Tiefen- und RGB-Bilder. Wie das bivariate Histogramm in der Abbildung 2.7b zeigt, ist die Anzahl der Observationen in den verschiedenen Klassen relativ ausbalanciert. Beim Einsatz der linearen SVMs werden dennoch zusätzlich klassenspezifische Wichtungsparameter einbezogen (siehe dazu Abschn. 4.2).

Die 931 RGB-und 2.5D-Bilder werden in dieser Arbeit als Trainings-, Validierungs-

und Testdaten zur Entwicklung des Feebackverfahrens eingesetzt. Ergänzend dazu dient ein zweiter Datensatz als finaler Testdatensatz. Dies bedeutet, dass er ausschließlich zur Evaluation des fertigen Verfahrens vorgesehen ist und zu keinem Zeitpunkt Teil des Entwicklungsprozesses war. Er umfasst jeweils 117 RGB- und 2.5D-Tiefenbilder von insgesamt fünf Patienten mit unterschiedlich stark ausgeprägter Fazialisparese und wurde im Rahmen von logopädischen Sitzungen in der Fachklinik Bad Liebenstein aufgenommen. Zwei Aufnahmen aus diesem Datensatz wurden bereits in der Abbildung 2.5 gezeigt, weitere Aufnahmen sind im Abschnitt 5.5.2 enthalten. Der finale Testdatensatz ist aus mehreren Gründen dazu geeignet, die Robustheit des entwickelten Feedbackverfahrens zu evaluieren:

- Er war nicht Teil des Entwicklungsprozesses.
- Er unterscheidet sich vom ersten Datensatz hinsichlich der Aufnahmeumgebung, der Beleuchtung, den dargestellten Personen und ihren Altersklassen.
- Er bildet ein reales Therapieszenario ab (Fazialisparese-Patienten, Schrägstellung des Kopfes in Folge der physischen Beeinträchtigung).

Bevor in den Kapiteln 3 bis 5 detaillierter auf einzelne Aspekte des Feedbackverfahrens eingegangen wird, wird im folgenden Abschnitt als Orientierungsrahmen ein Überblick über alle Komponenten der technischen Gesamtarchitektur gegeben.

2.4. Die technische Gesamtarchitektur

Der Fokus der nachfolgenden Kapitel 3 bis 5 liegt auf der Entwicklung und Evaluierung von einzelnen Teilkomponenten des in dieser Arbeit vorgestellten Feedbackverfahrens. Um die Einordnung dieser Teilaspekte in den Gesamtrahmen der Arbeit zu erleichtern, wird in diesem Unterkapitel ein Überblick über das Feedbackverfahren gegeben. Der Überblick orientiert sich dabei chronologisch am Aufbau der zu Grunde liegenden technischen Gesamtarchitektur. Sie ist in der Abbildung 2.8 gezeigt.

2.4.1. Vorverarbeitung

Die Kinect vereint zwei Kamerasysteme und nimmt sowohl RGB- als auch 2.5D-Tiefenbilder auf. Letzteres ist vergleichbar mit einem zweidimensionalen Intensitätsbild. Anstelle der Intensitätswerte enthält es jedoch Tiefenwerte, die die Distanz eines Objektpunktes zur Kamera beschreiben. Da der Farb- und der Tiefensensor unterschiedliche Abbildungseigenschaften aufweisen und räumlich verschoben sind, sind die

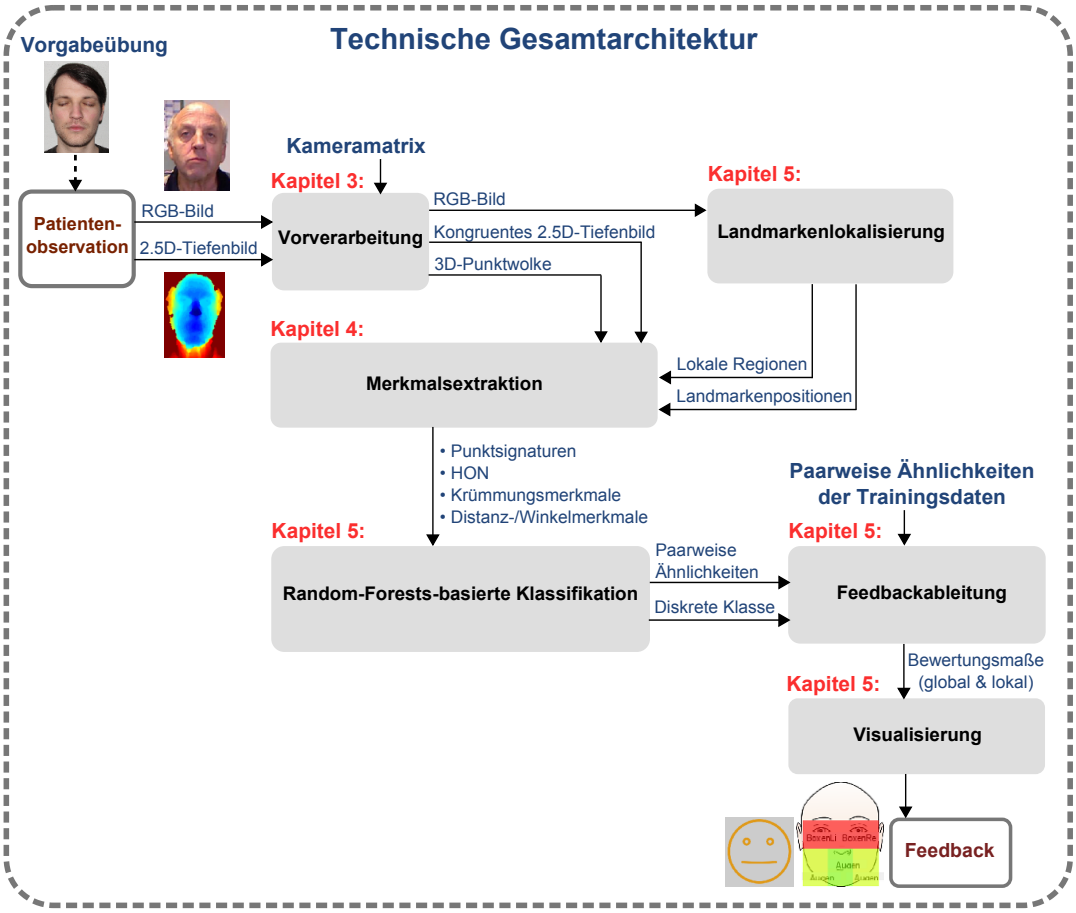
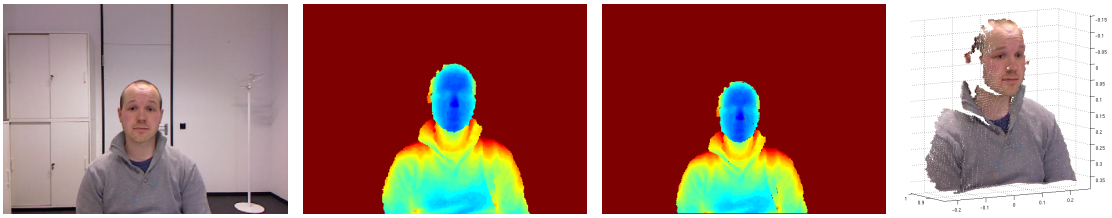


Abbildung 2.8.: Schematische Darstellung der technischen Gesamtarchitektur.

von der Kinect ausgegebenen RGB- und 2.5D-Tiefenbilder nicht deckungsgleich, wie die Abbildungen 2.9a und 2.9b verdeutlichen. Weil die Lokalisierung der Landmarken und Extraktionsareale in dieser Arbeit auf dem Farbbild erfolgt, die Merkmalsextraktion hingegen auf dem Tiefenbild, ist es erforderlich, die korrespondierenden Pixel beider Darstellungen zu ermitteln. Zu diesem Zweck soll das Tiefenbild so transformiert werden, dass es deckungsgleich zum RGB-Bild ist (vgl. Abb. 2.9c).

Für diese Transformation sind die sogenannten Kameramatrizen erforderlich, die im Rahmen der Kalibrierung ermittelt werden und die Eigenschaften der Aufnahmesysteme beschreiben. Die Kameramatrizen sind zudem vonnöten, um aus den Farb- und Tiefendaten eine dreidimensionale Punktwolke mit zugeordneten RGB-Farbwerten zu ermitteln (siehe Abb. 2.9d). Jeder Punkt einer Punktwolke repräsentiert eine Position in einem dreidimensionalen Koordinatensystem. Diese Darstellung hat gegenüber dem 2.5D-Bild einige Vorteile. So lassen sich euklidische Distanzen zwischen zwei 3D-Punkten berechnen, beispielsweise den Augeninnenwinkeln, die den tatsächlichen Abständen im Gesicht entsprechen. Auf Basis eines 2.5D-Bildes ist dies nicht mög-



(a) (b) (c) (d)

Abbildung 2.9.: **(a)** - **(b)** Zwei Beispiele für die von der Kinect ausgegebenen RGB- und 2.5D-Tiefenbilder. Die Tiefenwerte werden durch ein definiertes Farbschema repräsentiert, wobei der Verlauf von blau bis rot zunehmende Distanzwerte beschreibt. In der Darstellung ist die Farbskalierung so festgelegt, dass Tiefenwerte, die einen bestimmten Grenzwert übersteigen, dunkelrot eingefärbt sind. **(c)** Das transformierte 2.5D-Tiefenbild. **(d)** Die dreidimensionale Punktwolke. Jedem Punkt wurde der RGB-Wert des korrespondierenden Farbpixels zugeordnet.

lich, da die x- und y-Koordinaten lediglich die Position eines Pixels innerhalb einer zweidimensionalen Bildebene definieren. Zudem lässt ein Vergleich von zwei beliebigen 2.5D-Tiefenbildern kein Rückschluss auf relative Größenverhältnisse zu, da die Höhe und Breite eines aufgenommenen Gesichts in Abhängigkeit von der Aufnahmedistanz variiert.

Eine weitere Aufgabenstellung der Datenvorverarbeitung stellt die Korrektur ungültiger Tiefenwerte dar. Diese können zum Beispiel aufgrund bestimmter Materialeigenschaften, der Nichteinhaltung des für die Kinect empfohlenen Aufnahmeabstandes oder der gegenseitigen Verdeckung von Objekten entstehen. Da diese ungültigen Werte die nachfolgende Merkmalsextraktion beeinträchtigen, werden sie mit Hilfe einer in dieser Arbeit entwickelten Methode korrigiert.

Alle drei beschriebenen Aspekte der Vorverarbeitung, bestehend aus der Kalibrierung, der Punktwolkenerstellung und der Tiefenwertkorrektur, werden in Kapitel 3 beschrieben.

2.4.2. Landmarkenlokalisierung

Bevor aus den 2.5D-Tiefenbildern und den 3D-Punktwolken Merkmale extrahiert werden können, ist eine Lokalisierung und Eingrenzung von Extraktionsarealen notwendig. Im Rahmen dieser Arbeit erfolgt dies auf Basis von sogenannten Landmarken, die charakteristische Punkte innerhalb des Gesichts repräsentieren. Zu diesen zählen beispielsweise die Nasenspitze oder die Mundwinkel. Da ein manuelles Setzen dieser Landmarken den Übungsablauf beeinträchtigen würde, wird ein automatisiertes Ver-

fahren zur Landmarkenlokalisierung benötigt. Im Zuge der Umsetzung des Feedbackverfahrens wurden drei verschiedene Lokalisierungsansätze evaluiert. Die wesentlichen Ergebnisse sind im Abschnitt 5.5.1 zusammengefasst.

2.4.3. Merkmalsextraktionsverfahren

Nach der Vorverarbeitung der Tiefendaten und der Lokalisierung der Extraktionsareale erfolgt die eigentliche Analyse und Informationsgewinnung. Dazu werden Merkmalsdeskriptoren extrahiert, die eine niedrigdimensionale Repräsentation der Gesichtsdaten darstellen. Auf Grundlage einer systematischen Literaturrecherche in Unterkapitel 4.1 wurden fünf verschiedene Merkmalsextraktionsverfahren (Merkmalstypen) ausgewählt. Sie umfassen:

- Distanzmerkmale
- Winkelmerkmale
- Punktsignaturen
- Histogramme orientierter Normalenvektoren
- Krümmungsmerkmale

und werden in den Unterkapiteln 4.3 bis 4.6 vorgestellt und ausführlich evaluiert. Eine Übersicht über das zugrunde liegende Testszenario ist in Unterkapitel 4.2 gegeben. Für die folgenden Schritte des Feedbackverfahrens werden die extrahierten Merkmalsdeskriptoren aller fünf Merkmalstypen zu einem Vektor zusammengefasst und einem zuvor trainerten Random-Forest-Klassifikator übergeben.

2.4.4. Feedbackerzeugung

Die Feedbackerzeugung erstreckt sich innerhalb der technischen Gesamtarchitektur über zwei Komponenten und wird in den Unterkapiteln 5.1 bis 5.3 vorgestellt und evaluiert.

Die erste Komponente entspricht im Wesentlichen einer Klassifikation mit einem Random-Forest (RF), bei welcher der Patientenobservation eine von zwölf möglichen therapeutischen Übungsklassen zugeordnet wird (vgl. Abschn. 2.3.1). Ergänzend zur diskreten Klasse lassen sich die paarweisen Ähnlichkeiten zwischen der Patientenobservation und den Trainingsobservationen ermitteln. Diese beschreiben die prozentuale Häufigkeit mit der die Patientenobservation und eine bestimmte Trainingsobservation

in einem gemeinsamen Endknoten des RF landen. Sie dienen somit als Maß um die Ähnlichkeit zwischen zwei Merkmalsvektoren zu quantifizieren. Ist neben dem globalen auch lokales Feedback gewünscht, muss für jede Feedbackregion ein gesonderter RF auf Basis der lokal extrahierten Merkmalsdeskriptoren trainiert werden.

Innerhalb der zweiten Komponente werden die ermittelten paarweisen Ähnlichkeiten mit den trainingsdateninternen paarweisen Ähnlichkeiten verglichen. Diese lassen sich ebenfalls aus dem RF auslesen. Der Vergleich beschränkt sich jedoch ausschließlich auf die Trainingsobservationen der, gemäß dem Trainingsplan, auszuführenden Übung. Letztere wird in dieser Arbeit auch als Vorgabe- oder Zielübung bezeichnet. Das Ergebnis der Feedbackableitung ist ein Paar von zwei kontinuierlichen Bewertungsmaßen $\tilde{a}$ und $\tilde{v}$ (pro RF).

2.4.5. Visualisierung

Das globale Feedback bezieht sich auf das ganze Gesicht. Um eine höhere räumliche Auflösung der Übungsbewertung zu erreichen, werden in dieser Arbeit zusätzlich fünf lokale Feedbackregionen festgelegt. Im Ergebnis resultiert dies in sechs Paaren aus je zwei kontinuierlichen Bewertungsmaßen. Jedem Paar wird anschließend grenzwertbasiert eine von insgesamt sechs diskreten Feedbackstufen zugeordnet (siehe Abschn. 5.4). Um das erzeugte Feedback anschaulich demonstrieren zu können, wurde im Rahmen dieser Arbeit ein Prototyp mit grafischer Nutzeroberfläche implementiert. Der Prototyp umfasst insgesamt fünf Feedbackelemente, die im Unterkapitel 5.4 vorgestellt werden. Beispielergebnisse des globalen und regionenbezogenen Feedbacks werden im Unterkapitel 5.5 gezeigt und evaluiert.

2.5. Zusammenfassung

Der Fokus dieses Kapitels lag auf den konzeptionellen Aspekten der Arbeit. Zu Beginn wurden die Ergebnisse der anwendungsszenariobezogenen Literaturrecherche dokumentiert und eine Übersicht über die verschiedenen Ursachen und Erscheinungsformen von Mimikdysfunktionen gegeben. Danach erfolgte eine Beschreibung der beiden Datensätze, die den Entwicklungs- und Evaluierungsschritten zugrunde liegen. Abschließend wurde die technische Gesamtarchitektur des entwickelten Feedbackverfahrens in ihren Grundzügen beschrieben. Tiefergehende Informationen zur Vorverarbeitung der Daten, der Extraktion der Merkmalsdeskriptoren, sowie der Erzeugung und Visualisierung des Feedbacks finden sich in den Kapiteln 3 bis 5.

3. Aufnahme und Vorverarbeitung der Bilddaten

Neben den gewählten Merkmalsextraktions- und Feedbackverfahren trägt auch die Qualität und Form der RGB- und Tiefendaten zur Funktionsweise des Feedbacksystems bei. Aus diesem Grund liegt der Fokus dieses Kapitels auf der Aufnahme und Vorverarbeitung der Bilddaten. In den Unterkapiteln 3.1 und 3.2 finden sich einleitende theoretische und praktische Informationen zur Datenaufnahme und -verarbeitung. Diese bilden das Fundament für die folgenden Unterkapitel, in welchen die, im Rahmen dieser Arbeit durchgeführten, Vorverarbeitungsschritte dokumentiert sind. Eine detailliertere Gliederung dieses Kapitels ist in der Abbildung 3.1 gegeben. Die Einordnung der Vorverarbeitungskomponente in die technische Gesamtarchitektur wurde in der Abbildung 1.4 ersichtlich.

3.1. Theoretische Grundlagen des Kameramodells und der Kalibrierung

Zur Einführung in die Theorie und Notation der Kameraabbildung und -kalibrierung dient im Folgenden das einfachste theoretische Modell, das sogenannte *Lochkamera-*

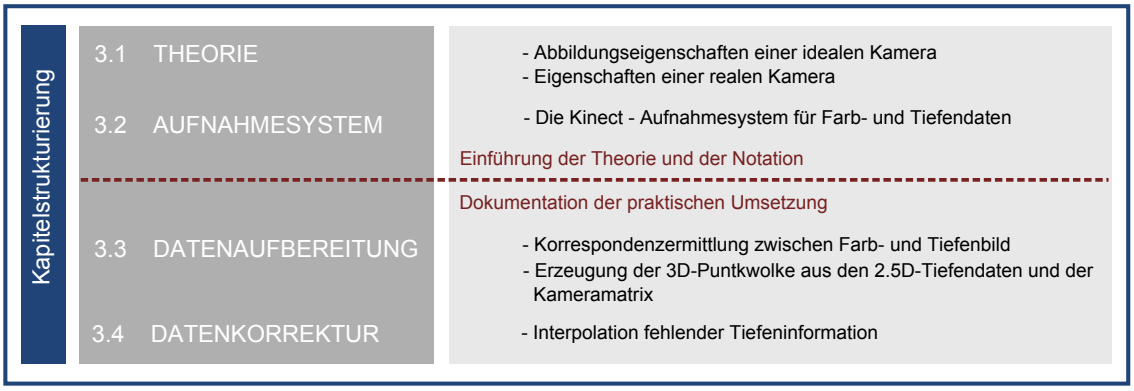


Abbildung 3.1.: Übersicht über die Gliederung dieses Kapitels.

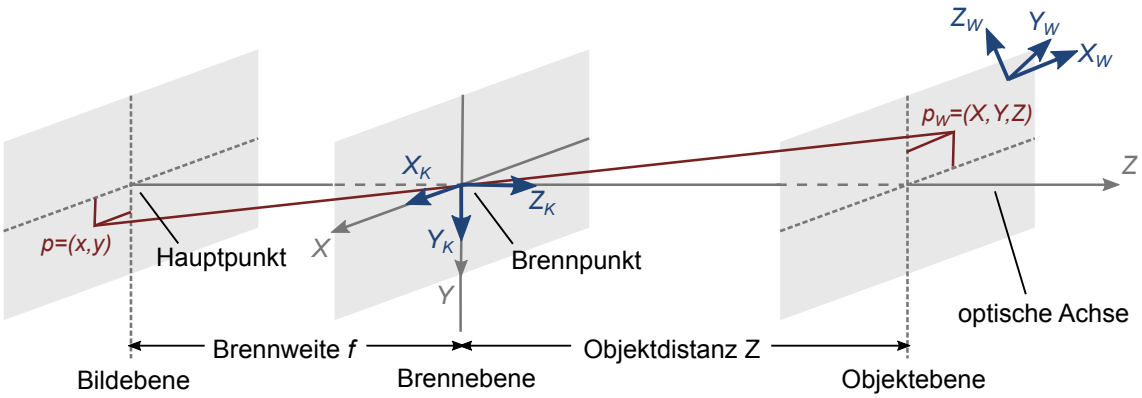


Abbildung 3.2.: Das theoretische Modell einer idealisierten Lochkamera. Abbildung in Anlehnung an [JÄHNE, 2002].

modell [JÄHNE, 2002]. Hierbei handelt es sich um ein idealisiertes Konstrukt, da die Abbildungseigenschaften von realen Kameras durch Eigenheiten der Bauweise und der Objektive beeinflusst werden, wie im Hauptteil dieses Unterkapitels deutlich wird.

3.1.1. Das Lochkameramodell

Das Lochkameramodell ist ein theoretisches Konstrukt zur Beschreibung der Funktionsweise einer Kamera. Aufgrund einer infinitesimal kleinen Lochblende in der Brennebene bildet es mit einem einzelnen Strahl einen dreidimensionalen Objektpunkt $\mathbf{p}_W = (X, Y, Z)^T$ auf einen Punkt $\mathbf{p} = (x, y)^T$ einer zweidimensionalen Ebene ab (siehe Abbildung 3.2). Die zweidimensionale Ebene bezeichnet man auch als Bildebene. Orthogonal zu Bild- und Brennebene verläuft die optische Achse. Sie durchstößt dabei Haupt- und Brennpunkt, welche im Abstand f voneinander entfernt liegen.

Die Beschreibung des dreidimensionalen Punktes $\mathbf{p}_W$ ist sowohl in Kamerakoordinaten $(X_K, Y_K, Z_K)^T$ als auch Weltkoordinaten $(X_W, Y_W, Z_W)^T$ möglich. Beim Kamerakoordinatensystem liegt der Ursprung im Brennpunkt der Kamera und die Z_K-Achse fällt mit der optischen Achse der Kamera zusammen. Das Weltkoordinatensystem orientiert sich am Objekt und ist nach [JÄHNE, 2002] so definiert, dass X_W und Y_W die horizontalen Richtungen beschreiben und Z_W die vertikale. Der Ursprung und die Ausrichtung des Weltkoordinatensystems ist prinzipiell beliebig, die Abbildung 3.2 zeigt nur eine beispielhafte Lage. Objekte können mittels Rotations- und Translationsoperationen von Weltkoordinaten in Kamerakoordinaten (und umgekehrt) überführt werden. Wenn nicht ausdrücklich anders erwähnt, sind in dieser Arbeit Kamera- und Weltkoordinatensystem identisch, d.h. das Weltkoordinatensystem hat seinen Ursprung im Brennpunkt und die Z_W-Achse verläuft entlang der optischen Achse. In diesem Fall entfallen die Überführungsoperationen zwischen den Koordinatensystemen

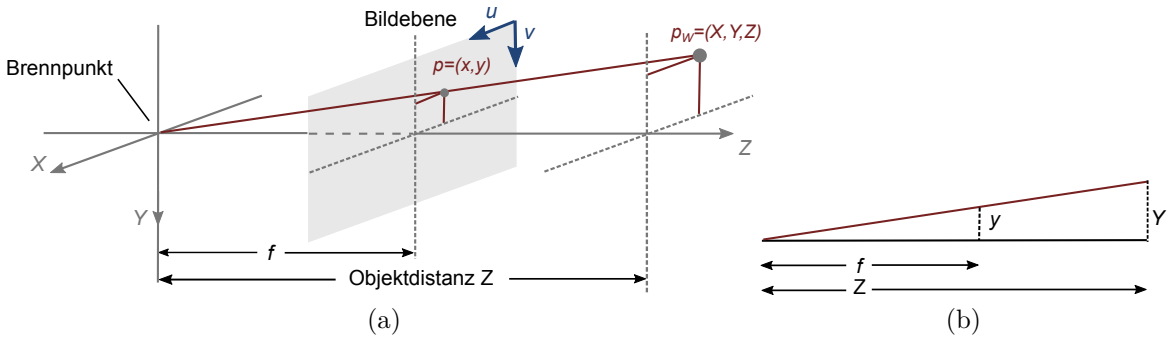


Abbildung 3.3.: **(a)** Die Bildebene wurde zur Vereinfachung der Berechnung zwischen den Brennpunkt und den 3D-Objektpunkt verschoben. Abbildung in Anlehnung an [Bradski und Kaehler, 2008]. **(b)** Darstellung der vereinfachten Strahlensatzberechnung.

und für $\mathbf{p}_W$ wird eine einfache Schreibweise ohne kennzeichnende Subskripte W oder K für die Koordinaten X, Y und Z gewählt.

Anhand der Strahlensätze lassen sich aus der Abbildung 3.2 folgende Beziehungen zwischen den 2D-Bild- und den 3D-Weltkoordinaten ermitteln [Bradski und Kaehler, 2008]:

$$x = -\frac{fX}{Z}, \quad y = -\frac{fY}{Z}. \tag{3.1}$$

Es wird deutlich, dass sich die Bildkoordinaten aus den mit dem Faktor $-\frac{f}{Z}$ skalierten Weltkoordinaten ergeben. Den Bildkoordinaten fehlt somit jegliche Information über absolute Größen und Entfernungen von Objekten [Jähne, 2002].

Zur Entfernung des negativen Vorzeichens in Gleichung 3.1 und Vereinfachung der Berechnung hat sich eine geänderte Darstellungsform durchgesetzt [Bradski und Kaehler, 2008]. Bei dieser wird die Bildebene zwischen den Brennpunkt und den Objektpunkt verschoben. Der betragsmäßige Abstand zwischen Haupt- und Brennpunkt bleibt dabei jedoch erhalten, wie die Grafiken in Abbildung 3.3 verdeutlichen.

Bei den bisherigen Betrachtungen ist das Koordinatensystem der Bildebene so definiert, dass der Ursprung im Bildmittelpunkt liegt. Üblicher ist jedoch die Positionierung des Ursprungs in der linken oberen Ecke der Bildebene. Die Berechnung der neuen Pixelkoordinaten (u, v) erfolgt dann über eine Addition mit den Koordinaten des Hauptpunktes $\mathbf{p}_0 = (p_x, p_y)$:

$$u = \frac{fX}{Z} + p_x, \quad v = \frac{fY}{Z} + p_y. \tag{3.2}$$

Hierbei werden p_x und p_y vom Ursprung des (u, v)-Koordinatensystems aus abgelesen.

3.1.2. Projektive Abbildungen

Bei der Abbildung von 3D-Weltpunkten auf 2D-Bildpunkte treten Phänomene auf, die aus der uns umgebenden dreidimensionalen Welt nicht bekannt sind. So erhalten parallele Geraden einen gemeinsamen Schnittpunkt im Unendlichen, den sogenannten Fluchtpunkt [Rahmann und Burkhardt, 2011]. Dieses Phänomen tritt beispielsweise bei Aufnahmen von in Richtung Horizont verlaufenden Straßen oder Häuserschluchten auf. Allein über euklidische Transformationen lässt sich dies nicht beschreiben, weshalb wir auf eine erweiterte Form der Geometrie zurückgreifen, die sogenannten *projektiven Transformationen* bzw. *projektiven Abbildungen*. Diese machen von *homogenen Koordinaten* Gebrauch, welche durch eine Tilde gekennzeichnet werden. Ein Punkt im $\mathbb{R}^n$ wird in euklidischen Koordinaten als n-dimensionaler Vektor repräsentiert, während die Darstellung über homogene Koordinaten einem $(n+1)$-dimensionalen Vektor entspricht. Für einen zweidimensionalen Punkt $\mathbf{p} = (x, y)^T$ in einer Bildebene ergibt sich ein dreidimensionaler Vektor $\tilde{\mathbf{p}} = (x, y,1)^T$. Homogene Koordinaten sind invariant gegenüber Skalierungen, weshalb für alle $\lambda \neq 0$ gilt:

$$\tilde{\mathbf{p}} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \sim \begin{pmatrix} \lambda x \\ \lambda y \\ \lambda \end{pmatrix}, \ \lambda \neq 0, \ \lambda \in \mathbb{R}. \tag{3.3}$$

Die durch Gleichung 3.3 definierten Punkte werden nach [Hartley und Zisserman, 2003] als finite Punkte bezeichnet, homogene Vektoren mit dem Wert 0 in der (n+1)-ten Koordinate als ideale Punkte beziehungsweise Punkte im Unendlichen. Der Nullvektor für $\lambda = 0$ ist nicht definiert [Rahmann und Burkhardt, 2011].

Der Vorteil der homogenen Koordinaten ist, dass sie die Berechnung des Abbildungsprozesses erleichtern. Durch die Zusammenfassung der Hauptpunktkoordinaten (p_x, p_y) und der Brennweite f in einer Kameramatrix $\mathbf{K}$:

$$\mathbf{K} = \begin{pmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.4}$$

wird für die Berechnungen in Gleichung 3.2 eine Matrixmultiplikation ermöglicht:

$$\tilde{\mathbf{p}} = \begin{pmatrix} \lambda u \\ \lambda v \\ \lambda \end{pmatrix} = \begin{pmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{3.5}$$

Durch Division von $\tilde{\mathbf{p}}$ mit λ und dem Entfernen der dritten Dimension, ergeben sich

die zweidimensionalen euklidischen Bildkoordinaten $\mathbf{p} = (u, v)^T$.

3.1.3. Erweiterung des idealen Modells

Bei der vorgestellten Lochkamera handelt es sich um ein idealisiertes, theoretisches Modell. Die Abbildungseigenschaften von realen Kameras weichen, aufgrund verschiedener Einflussfaktoren, teilweise deutlich von denen des idealisierten Modells ab. Drei dieser Faktoren werden in den folgenden Unterabschnitten näher beleuchtet. Sie umfassen:

- die Verschiebung des Hauptpunktes in x- und y-Richtung
- Verzeichnungen des Kamerabildes
- das nicht-quadratische Seitenverhältnis der Bildsensorpixel

Die Unterabschnitte dienen der allgemeinen theoretischen Übersicht und gehen nicht näher auf das in dieser Arbeit verwendete Kamerasystem ein. Eine detaillierte Analyse der Kinect und ihrer technischen Daten findet sich in Abschnitt 3.2.

Verschiebung des Hauptpunktes

Bei handelsüblichen Kameras liegen die Werte der Bildsensorhöhe und -breite im Millimeterbereich. Abhängig von der Bildauflösung ergeben sich für die darauf befindlichen einzelnen Pixel Seitenmaße von wenigen Mikrometern Länge. Bei der Herstellung der Kameras ist aufgrund dieser kleinen Ausmaße nicht immer gewährleistet, dass die optische Achse den Bildsensor exakt in der Mitte durchstößt [Bradski und Kaehler, 2008]. Diese Verschiebung des Hauptpunktes sowohl in x- als auch y-Richtung hat Auswirkungen auf den Abbildungsprozess und muss in entsprechende Berechnungen einbezogen werden. Die Werte dieser Verschiebungen werden bei der Kamerakalibrierung ermittelt. Die im Abschnitt 3.1.1 eingeführten Parameter p_x und p_y entsprechen daher bei realen Kamerasystemen nicht zwingend der Position der exakten Bildmitte.

Verzeichnungen des Bildes

Die Abbildung eines 3D-Weltpunktes auf eine 2D-Bildebene übersteigt die dem Menschen im Alltag geläufige euklidische Geometrie von Translation und Rotation, wie das im vorhergehenden Abschnitt beschriebene Phänomen der parallelen Geraden mit virtuellem Schnittpunkt zeigt. Neben diesem dem Abbildungsprozess innewohnenden Effekt, können auch konstruktionsbedingte Eigenschaften der Kamera Auswirkungen

auf die Geometrie der Abbildung haben und zu Verzerrungen des Bildes führen. Die bekanntesten Formen sind die radiale und die tangentiale Verzeichnung.

Die *radiale Verzeichnung* entsteht durch die uneinheitliche Lichtbeugung innerhalb einer Linse. Während Lichtstrahlen entlang der optischen Achse unbeeinflusst bleiben, nimmt ihre Beugung mit wachsendem radialen Abstand zum Linsenzentrum zu. Die entstehenden Verzeichnungen sind kissen- oder tonnenförmig und gerade Linien werden gebogen dargestellt. In der Literatur finden sich verschiedene Modelle zur Beschreibung der radialen Verzeichnung $D_R(r)$. Eine geläufige Variante ist die Annäherung über die ersten drei Glieder einer Taylorreihe ([Bradski und Kaehler, 2008], [Pears et al., 2012]):

$$\begin{aligned} x_d &= x \cdot D_R(r) = x \cdot \left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right), \\ y_d &= y \cdot D_R(r) = y \cdot \left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right). \end{aligned} \tag{3.6}$$

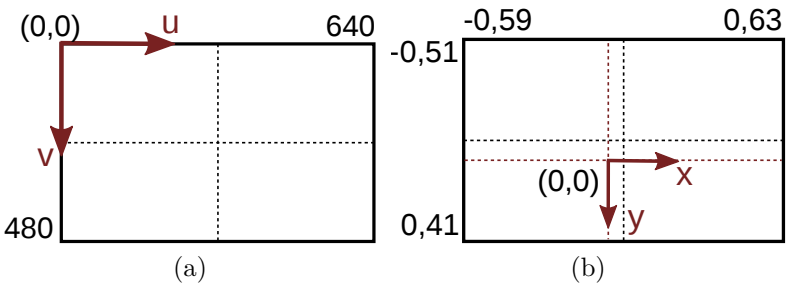
Dabei entspricht $(x, y)^T$ den Koordinaten der vom idealen Lochkamera-Modell prädizierten Position und $(x_d, y_d)^T$ gibt die Position nach Einwirken der radialen Verzeichnung an. Der Parameter r beschreibt den Radius des betrachteten idealen Punktes zum Linsenzentrum mit $r^2 = x^2 + y^2$.

In Gleichung 3.6 ist ersichtlich, dass bei der Berechnung der radialen Verzeichnung mehrfache Potenzen des Radius verwendet werden und $D_R(r)$ einem variablen Skalierungsfaktor gleicht, der sich in Standardfällen um den Wert 1 bewegt. Dabei ergibt $D_R(r) < 1$ tonnenförmige und $D_R(r) > 1$ kissenförmige Verzeichnungen. Weil die radiale Verzeichnung von der Bildmitte ausgeht und unabhängig von der Bildgröße in Pixeln ist, werden die Pixelkoordinaten (u, v) *vor* Anwendung der Gleichung 3.6 mit der Brennweite normalisiert und in den Hauptpunkt verschoben. Aufgrund der projektiven Koordinaten ist dies über die Multiplikation mit der Inversen der Kameramatrix möglich. Das Resultat sind sogenannte *verallgemeinerte Bildkoordinaten*:

$$\tilde{\mathbf{p}} = \begin{pmatrix} \lambda x \\ \lambda y \\ \lambda \end{pmatrix} = \mathbf{K}^{-1} \cdot \begin{pmatrix} u \\ v \\ 1 \end{pmatrix}. \tag{3.7}$$

Die Abbildung 3.4 verdeutlicht, welchen Einfluss die Multiplikation der Pixelkoordinaten mit der Inversen der Kameramatrix $\mathbf{K}$ auf die Positionierung und den Wertebereich des resultierenden Koordinatensystems hat. Es wurde beispielhaft eine Kinectaufnahme mit den Ausmaßen 640×480 Pixeln zur Berechnung eingesetzt. Der Hauptpunkt liegt, vom Ausgangskoordinatensystem aus gesehen, in $(u = 308, v = 266)^T$.

Für Standardobjektive sind die Parameter k_1 und k_2 zur Modellierung der radia-



(a) (b)

Abbildung 3.4.: **(a)** Das Ausgangskoordinatensystem mit dem Ursprung in der linken oberen Bildecke. Die Bildbreite entspricht 640 Pixeln entlang der u-Achse und 480 Pixeln entlang der v-Achse. **(b)** Nach der Multiplikation mit der Inversen der Kameramatrix befindet sich der Ursprung des Koordinatensystems im Hauptpunkt und der Wertebereich der Achsen ist verkleinert.

len Verzeichnung von vorrangiger Bedeutung [Bradski und Kaehler, 2008]. Der Parameter k_3 kann eingebunden werden, wenn sehr starke Verzeichnungen, wie sie beispielsweise bei Fischaugen-Objektiven auftreten, beschrieben werden sollen. Die Abbildung 3.5 illustriert verschiedene Formen der radialen Verzeichnung, angewandt auf ein künstlich erzeugtes Schachbrettmuster. Es ist erkennbar, dass die Verzeichnung im Bereich des Bildzentrums am geringsten ist. Im Hauptpunkt ($r = 0$) existiert keine Verzeichnung, da $D_R(r) = 1$.

Die *tangentiale Verzeichnung* entsteht, wenn die Linse nicht parallel zur Ebene des Bildsensors ausgerichtet ist. Die Berechnung der verzerrten Koordinaten erfolgt über [Bradski und Kaehler, 2008]:

$$\begin{aligned} x_d &= x + D_{Tx}(x, y, r) = x + [2p_1xy + p_2\left(r^2 + 2x^2\right)] \\ y_d &= y + D_{Ty}(x, y, r) = y + [2p_2xy + p_1\left(r^2 + 2y^2\right)]. \end{aligned} \tag{3.8}$$

Die Abbildung 3.6 zeigt mehrere Beispiele für unterschiedliche Belegungen der Parameter p_1 und p_2.

Eine Kombination beider Verzeichnungen ergibt:

$$\begin{aligned} x_d &= x \cdot D_R(r) + D_{Tx}(x, y, r) \\ y_d &= y \cdot D_R(r) + D_{Ty}(x, y, r). \end{aligned} \tag{3.9}$$

In der Gleichung 3.9 werden verzerrte Koordinaten (x_d, y_d) aus idealen Lochkamerakoordinaten ermittelt. Für praktische Anwendungen ist hingegen der umgekehrte Weg von Interesse, da Kameras verzerrte Koordinaten ausgeben, jedoch ideale gewünscht sind. Die Abhängigkeit der Verzerrungsterme von den Variablen x, y und r

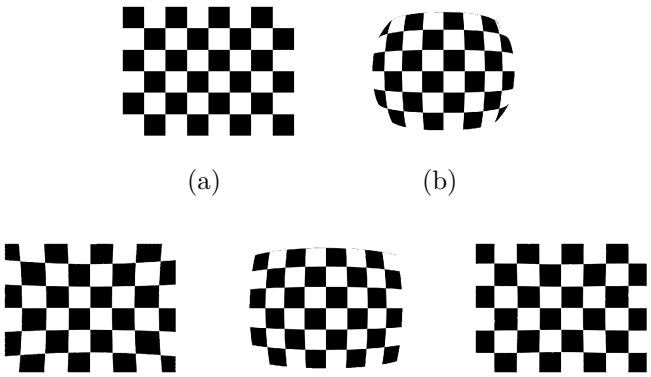


Abbildung 3.5.: Beispiele für die Anwendung der radialen Verzeichnung auf ein künstlich erzeugtes ideales Schachbrettmuster. **(a)** Ausgangsmuster **(b)** Die von Fischaugen-Objektiven bekannte Verzerrung entsteht, wenn alle drei Verzeichnungsparameter belegt sind und die negativen Werte dominieren (im Beispiel: $k_1 = -0{,}3$, $k_2 = -0{,}3$, $k_3 = -0{,}3$). Zur Beschreibung der Verzeichnung von Standardobjektiven ist der Wert k_3 von geringerer Bedeutung, weshalb er bei den weiteren Beispielen den Wert 0 zugewiesen bekommt. **(c)** Die sogenannte kissenförmige Verzeichnung (*engl.* pincushion distortion) entsteht beispielsweise bei Werten von $k_1 = +0{,}3$, $k_2 = 0$. **(d)** Die tonnenförmige Verzeichnung (*engl.* barrel distortion) im Beispiel wird durch die Parameterwerte $k_1 = -0{,}3$, $k_2 = 0$ erzeugt. **(e)** Eine Mischung der kissen- und tonnenförmigen Verzeichnung ist ebenfalls möglich ($k_1 = 0{,}3$, $k_2 = -0{,}5$).

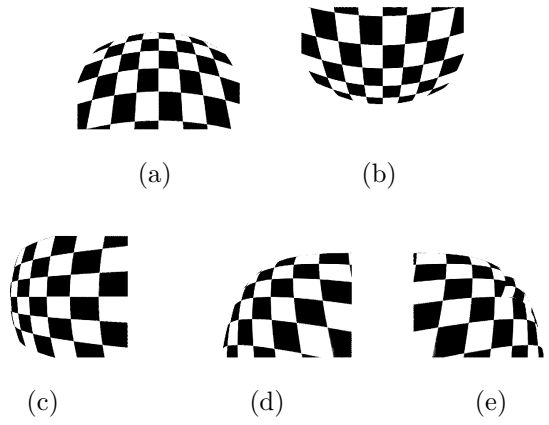


Abbildung 3.6.: Beispiele für die tangentiale Verzeichnung. **(a)** $p_1 = 0{,}3$, $p_2 = 0$ **(b)** $p_1 = -0{,}3$, $p_2 = 0$ **(c)** $p_1 = 0$, $p_2 = 0{,}3$ **(d)** $p_1 = 0{,}3$, $p_2 = 0{,}3$ **(e)** $p_1 = 0{,}3$, $p_2 = -0{,}3$.

führt allerdings dazu, dass die Berechnungen nicht invertierbar sind, weil die Werte der Variablen nicht rekonstruiert werden können. Da jedoch keine Abhängigkeit des Verzerrungsterms zum Bildinhalt besteht, ist es möglich, zuerst den Weg *ideale Koordinaten* → *verzerrte Koordinaten* zu berechnen und dann die gewonnenen Zusammenhänge in einer Lookup-Tabelle zu speichern, um diese für den umgekehrten Vorgang auszulesen und zu interpolieren.

Seitenverhältnis der Bildsensorpixel

Bei vielen Kamerasensoren sind die Bildsensorpixel nicht quadratisch [Bradski und Kaehler, 2008]. Zur Korrektur werden die Skalierungsfaktoren s_x und s_y eingesetzt. Sie beschreiben die Anzahl der Pixel pro festgelegter Einheitsdistanz. Korrigiert werden die Hauptpunkte p_x und p_y, sowie die Brennweite f. Aufgrund der gesonderten Skalierungsfaktoren für die Höhe und die Breite, entstehen zwei korrigierte Brennweiten $f_x = fs_x$ und $f_y = fs_y$. Die korrigierten Hauptpunktkoordinaten ergeben sich ebenfalls durch Multiplikation mit den Skalierungsfaktoren: $x_0 = p_x \cdot s_x$ und $y_0 = p_y \cdot s_y$.

3.1.4. Intrinsische und extrinsische Kameraparameter

Die durch die Kameramatrix beschriebenen Parameter f_x, f_y, x_0 und y_0 beeinflussen, wie auch die radialen und tangentialen Verzeichnungsparameter, die Abbildung eines 3D-Weltpunktes auf die 2D-Bildebene. Sie alle sind charakteristische, kamerainterne Kennwerte und werden unter dem Begriff der *intrinsischen Parameter* zusammengefasst [Bradski und Kaehler, 2008]. Sie beschreiben die Abbildungseigenschaften *einer* Kamera. Ergänzend dazu stehen die *extrinsischen Parameter*, welche die relative Position einer Kamera zu einem anderen System definieren [Bradski und Kaehler, 2008]. Bei diesem System kann es sich um einen Objektpunkt oder um eine weitere Kamera handeln. Die Kinect – bestehend aus einer Farb- und einer Tiefenkamera – ist ein solches Stereosystem, bei dem sich beide Kameras in starrer, das heißt zueinander unveränderlicher Lage, befinden.

Um korrespondierende Pixel zweier Kameras zu ermitteln und ihre Information (z.B. Farb- und Tiefeninformation) miteinander kombinieren zu können, muss die relative Position zwischen diesen Kameras bekannt sein. Sie wird über eine Rotationsmatrix $\mathbf{R}$ und einen Translationsvektor $\mathbf{t}$ angegeben. Die Überführung eines Punktes $\mathbf{p}_{WL}$, dessen Position im Koordinatensystem der linken Kamera beschrieben ist, in einen Punkt $\mathbf{p}_{WR}$, mit Verortung im Koordinatensytem der rechten Kamera, ergibt sich aus:

$$\mathbf{p}_{WR} = \mathbf{R} \cdot \mathbf{p}_{WL} + \mathbf{t}. \tag{3.10}$$

Sowohl die intrinsischen als auch die extrinsischen Parameter können über die *Kalibrierung der Kamera* geschätzt werden. Dazu benötigt man Korrespondenzen von mindestens sechs Weltpunkten und ihren entsprechenden Vertretern in der 2D-Bildebene [Hartley und Zisserman, 2003]. Die Verwendung eines Kalibriermusters erleichtert die Detektion und Zuordnung dieser korrespondierenden Punkte, häufig wird hierbei ein planares Muster eingesetzt [Zhang, 2000]. Detaillierte Betrachtungen zur Schätzung der intrinsischen und extrinsischen Kameraparameter aus den Punktkorrespondenzen übersteigen den Rahmen dieser Arbeit. Für weiterführende Informationen sei auf [Hartley und Zisserman, 2003] verwiesen. Die praktische Durchführung der Kalibrierung des in dieser Arbeit verwendeten Kinectsystems ist in Abschnitt 3.3 dokumentiert.

3.2. Die Kinect

Das Aufnahmesystem der Wahl ist die Kinect für die Xbox 360 von Microsoft [MICROSOFT-XBox]. Sie ist in der Lage, sowohl Farb- als auch Tiefendaten aufzunehmen. Die Technologie zur Tiefenmessung wurde von der israelischen Firma PrimeSense entwickelt [Menna et al., 2011]. Die folgenden Abschnitte beschreiben Eigenschaften und Messprinzip der Kinect.

3.2.1. Eigenschaften

Die Kinect besteht aus einem Infrarotprojektor, einer Infrarotkamera und einer RGB-Kamera, die in fester Position zueinander in einem Gehäuse angeordnet sind. Für den Infrarotprojektor und die -kamera werden im Folgenden auch die Abkürzungen IR-Projektor und IR-Kamera verwendet. Neben den visuellen Bauteilen, enthält die Kinect noch ein Mikrofon, welches im Rahmen dieser Arbeit allerdings keinen Einsatz findet. Die Abbildung 3.7 zeigt die Elemente der Kinect und Beispiele für die von ihr ausgegebenen Bilddaten: ein RGB-, ein Infrarot- und ein 2.5D-Tiefenbild. Das 2.5D-Bild besteht, ähnlich wie ein einkanaliges Grauwertbild, aus zwei Dimensionen, enhält pro Pixel aber einen Tiefenwert anstelle eines Intensitätswertes. Der Tiefenwert beschreibt die Distanz der aufgenommenen Objekte zur Kamera in Form von normalisierten Disparitäten. Diese werden als Integerzahlen in einen Wertebereich von 0 bis 2047 codiert (11 bit). Das Infrarotbild dient hauptsächlich der kinect-internen Erzeugung des Disparitätsbildes und ist zur Verwendung als Infrarotbild selbst nicht

(a)

(b)

(c)

(d)

Abbildung 3.7.: **(a)** Der Kinect-Sensor und seine Elemente (Bildquelle der Kinect-Fotografie: Wikimedia Commons). **(b)** - **(d)** Die folgenden Abbildungen wurden mittels einer ASUS-Kamera aufgenommen, stimmen aber mit den Ausgabedaten einer Kinect überein: (b) RGB-Bild. (c) Korrespondierendes 2.5D-Tiefenbild. Die Distanzen sind farbcodiert, wobei der Farbverlauf blau-orange einen zunehmenden Kamerabstand beschreibt (rotbraun: ungültige Tiefenwerte). (d) Infrarot-Bild mit überlagertem Speckle-Muster.

geeignet, da es mit einem Speckle-Muster überlagert ist. Auf den Zweck des Speckle-Musters wird in diesem Abschnitt einige Absätze weiter unten eingegangen.

Sowohl die RGB- als auch die IR-Sensoren haben eine Auflösung von 1280×1024 Pixeln. Aufgrund der Bandbreitenbegrenzung der USB-Verbindung liegen das berechnete Disparitätsbild und das gestreamte RGB-Bild allerdings nur in einer Auflösung von 640×480 Pixeln vor. Die Übertragungsrate beträgt 30 Bilder pro Sekunde. Die Reichweite der Kamera bewegt sich in einem Aufnahmebereich von 0.5 bis 5 Metern, wobei der Fehler der Tiefeninformation proportional zum Quadrat der Distanz zur Kamera ist [Khoshelham und Elberink, 2012]. Zusätzlich nimmt die Dichte der Tiefeninformation ab, weil bei größerer Entfernung zur Kamera (und fehlendem Zoom) die Abbildung des Gesichts durch weniger Pixel erfolgt als bei geringerer Distanz. Für das Übungsszenario wird daher eine Personen-Kamera-Distanz von 0.5 bis 1.2 Metern festgelegt.

Die Kinect wurde ursprünglich als neuartige Mensch-Maschine-Schnittstelle für die Videospielkonsole Xbox 360 entwickelt und im November 2010 veröffentlicht. Das neue Konzept weckte schnell das Interesse verschiedener Programmierer, weshalb kurz nach

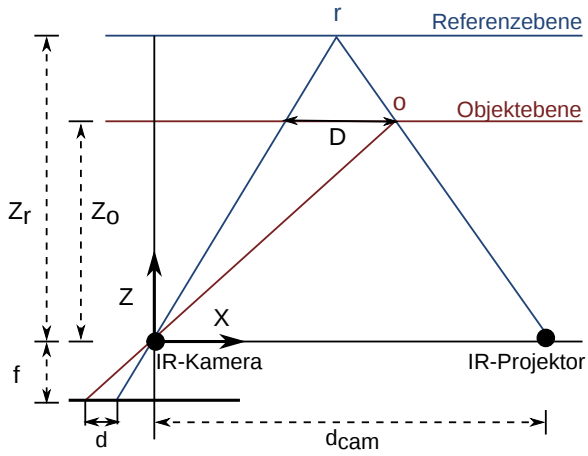


Abbildung 3.8.: Triangulation zur Ermittlung der Disparitäten (Abbildung nach [Khoshelham und Elberink, 2012]).

ihrer Markteinführung open-source Treiber und Softwareentwicklungsumgebungen wie das OpenNI-Framework Verbreitung fanden. Im Juni 2011 veröffentlichte Microsoft selbst ein SDK (*Software-Development Kit*) für die nichtkommerzielle Nutzung. Im Rahmen dieser Arbeit wird ein von Dirk-Jan Kroon entwickelter OpenNI-Wrapper für Matlab zum Stream der Farb- und 2.5D-Bilder eingesetzt. Er ist im *File Exchange* auf der Seite *www.mathworks.com* verfügbar[1]. Im folgenden Abschnitt wird eine kurze Übersicht über das Tiefenmessprinzip der Kinect gegeben.

3.2.2. Messprinzip

Die Tiefenmessung erfolgt durch das Prinzip der Triangulation [Khoshelham und Elberink, 2012]. Der Infrarotprojektor emittiert einen einzelnen Infrarotstrahl ($\lambda = 830\,nm$), der anhand eines optischen Gitters in ein Speckle-Muster gespalten wird und die zu vermessende Szenerie beleuchtet. Die Infrarotkamera erfasst das entstandene Speckle-Muster. Anschließend wird die Korrelation dieses Musters zu verschiedenen Referenzmustern ermittelt. Die Referenzmuster liegen im Speicher der Kinect vor und sind Referenzebenen mit bekanntem Kameraabstand zugeordnet. Ein einzelner Speckle-Punkt verschiebt sich in horizontaler Richtung, je nachdem, ob sich der Messpunkt vor oder hinter einer Referenzebene befindet. Diese Verschiebungen werden gemessen, um das Disparitätsbild zu erzeugen. Anhand der Disparität kann für jeden Objektpunkt die metrische Distanz zum Sensor berechnet werden.

Um die auf der Triangulation basierende Abstandsberechnung eines Objektpunktes o zu visualisieren, betrachten wir eine schematische Darstellung der Kinect (siehe Ab-

[1] http://www.mathworks.com/matlabcentral/fileexchange/30242-kinect-matlab, letzter Abruf: 06.02.2014

bildung 3.8). Diese umfasst einen Infrarotprojektor und eine Infrarotkamera, welche in unveränderlicher relativer Lage und einem Abstand von d_{cam} zueinander angeordnet sind. Gesucht wird die Distanz Z_o des Objektpunktes o zum Brennpunkt der IR-Kamera. Im diesen Brennpunkt wird der Ursprung eines 3D-Koordinatensystems gelegt. Die x-Achse des Koordinatensystems verläuft entlang der Verbindungslinie zwischen IR-Kamera und Infrarotprojektor. Die z-Achse liegt orthogonal dazu in Richtung der optischen Achse der IR-Kamera und die y-Achse sitzt orthogonal auf der x-z-Ebene auf. Vor der Kamera befindet sich, parallel zur x-Achse, eine beispielhafte Referenzebene, deren Abstand einem Referenzabstand Z_r aus dem Kinect-Speicher entspricht. Die blaue Darstellung visualisiert den Verlauf eines vom IR-Projektor ausgesandten Speckle-Punktes, welcher auf ein auf der Referenzebene befindliches Objekt r auftrifft und anschließend auf der Bildebene der IR-Kamera abgebildet wird. Die Bildebene befindet sich im Abstand der Brennweite f zum Kamerabrennpunkt. Wird nun vor der Referenzebene eine Objektebene mit einem Punkt o positioniert, wird der Speckle-Strahl bereits im Punkt o abgelenkt und fällt - entlang der x-Achse um die Disparität d verschoben - ebenfalls auf die Bildebene der IR-Kamera. Mit Hilfe der Strahlensätze lassen sich folgende Verhältnisse ermitteln [Khoshelham und Elberink, 2012]:

$$\frac{D}{d_{cam}} = \frac{Z_r - Z_o}{Z_r}, \tag{3.11}$$

und

$$\frac{d}{f} = \frac{D}{Z_o}. \tag{3.12}$$

Die konstanten Parameter f, d_{cam} und Z_r sind werkseitig bekannt und die unbekannte Verschiebung D wird unter Kombination der Gleichungen 3.11 und 3.12 entfernt. Durch Umstellen nach Z_o ergibt sich folgendes mathematisches Model zur Ermittlung der Tiefendistanz:

$$Z_o = \frac{Z_r}{1 + \frac{dZ_r}{f d_{cam}}}. \tag{3.13}$$

Für jedes Pixel des 2.5D-Bildes wird, unter Verwendung von Gleichung 3.13, ein Tiefenwert Z_o ermittelt [Khoshelham und Elberink, 2012].

3.3. Aufbereitung der Bilddaten

Die Farb- und die Infrarotkamera der Kinect sind räumlich versetzt angeordnet und projizieren die Szeneninformation mit unterschiedlichen Brennweiten auf den Bild-

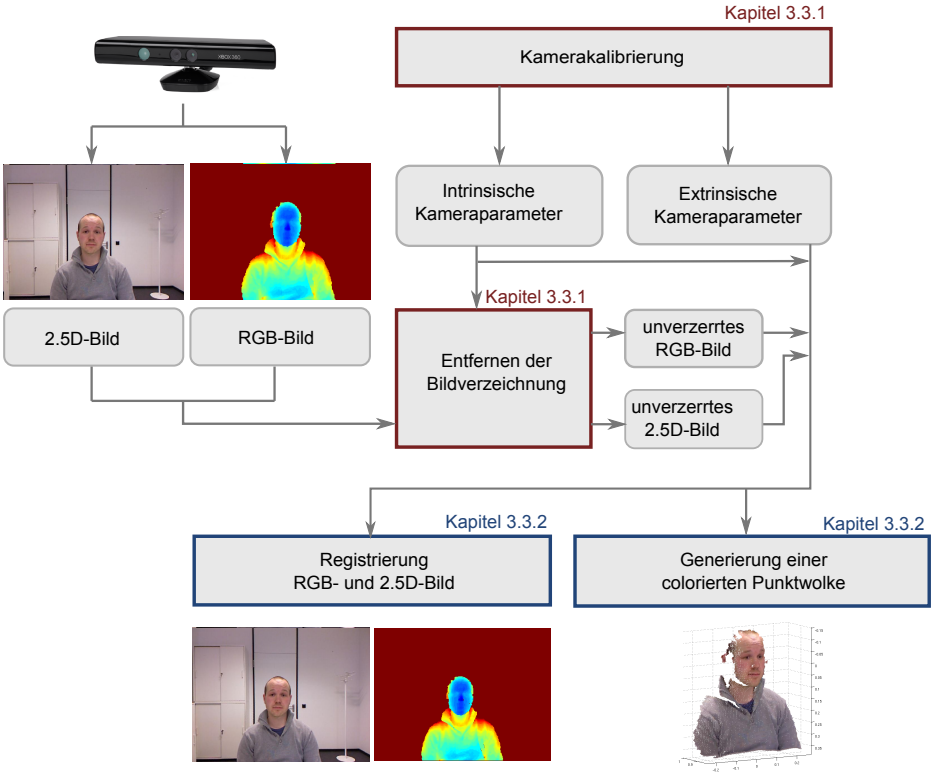
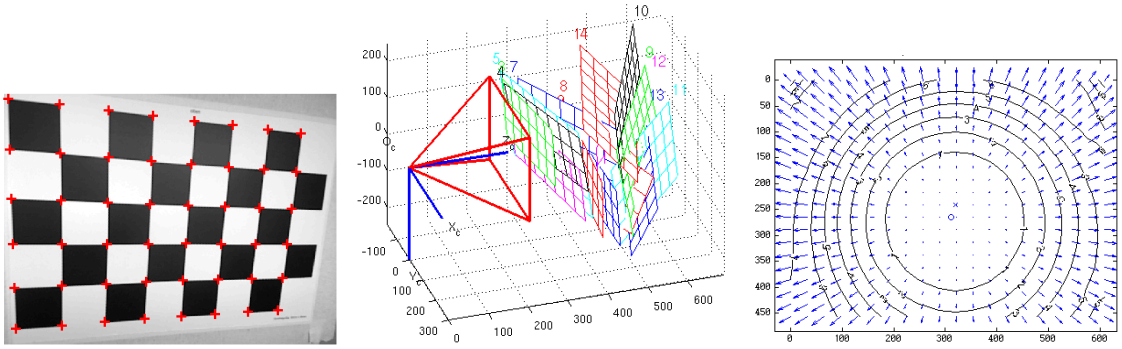


Abbildung 3.9.: Arbeitsschritte zur Registrierung der Farb- und Tiefeninformation, sowie zur Erzeugung der Punktwolken. Rechtecke mit abgerundeten Ecken symbolisieren Daten, Rechtecke mit spitzen Ecken Arbeitsschritte.

sensor. Dies hat zur Folge, dass Farb- und Tiefenbild nicht deckungsgleich sind und die Größe eines abgebildeten Objektes in beiden voneinander abweicht. Um bei der nachfolgenden Bildanalyse aus einer Vielfalt von Informationen schöpfen zu können, sollen beide Darstellungen miteinander registriert werden. Dazu wird das Tiefenbild so transformiert, dass es deckungsgleich zum Farbbild ist und korrespondierende Tiefen- und Farbwerte einander zugeordnet werden können. In einem weiteren Schritt werden Punktwolken erzeugt, welche die Darstellung der Tiefeninformationen auf einen dreidimensionalen, metrischen Raum ausdehnen. Die Grundlage für diese Schritte bilden die bei der Kamerakalibrierung gewonnenen Parameter, weshalb im Folgenden näher auf die praktische Umsetzung der Kalibrierung und die Entfernung der Bildverzeichnung eingegangen wird. Die Abbildung 3.9 visualisiert die Einzelschritte des in den folgenden Abschnitten beschriebenen Ablaufs.



(a) (b) (c)

Abbildung 3.10.: **(a)** Kalibriermuster mit automatisiert lokalisierten Markern. **(b)** Aufnahmepositionen des Kalibriermusters zur Kalibrierung der RGB-Kamera. Das Koordinatensystem ist kamerazentriert (Abbildung erstellt mittels *Camera Calibration Toolbox for Matlab*, Nr.5 entspricht (a)). **(c)** Modell der geschätzten kombinierten radialen und tangentialen Verzeichnung (RGB-Bild).

3.3.1. Praktische Umsetzung der Kamerakalibrierung

Um die Abbildungseigenschaften einer Kamera mittels Kalibrierung schätzen zu können, benötigt man korrespondierende 3D-Objekt- und 2D-Bildpunkte. Da die manuelle Lokalisierung dieser Punkte in den aufgenommenen Bilddaten zeitintensiv und ungenau sein kann, haben sich Verfahren zur automatisierten Detektion etabliert. Die automatisierte Detektion wird durch die Verwendung eines planaren Schachbrett-Kalibriermusters erleichtert, welches aus 8×5 Quadraten (Seitenlänge je 5 cm) besteht. Die Ecken der Schachbrettquadrate dienen als Positionen für die korrespondierenden Punkte und können unter Verwendung des Harris-Eckendetektors robust lokalisiert werden [Harris und Stephens, 1988], wie die Abbildung 3.10a zeigt. Ein entsprechendes Skript findet sich in der *Camera Calibration Toolbox for Matlab* von Jean-Yves Bouguet[2].

Die Toolbox ermöglicht zudem, basierend auf den gefundenen Punktkorrespondenzen, die Bestimmung der intrinsischen und extrinsischen Parameter. Da der Grauwertverlauf des Schachbretts im Tiefenbild nicht sichtbar und somit keine Eckpunktlokalisierung möglich ist, wird stattdessen das Infrarotbild verwendet, welches die Berechnungsgrundlage der Tiefeninformation bildet. Da die im Infrarotbild sichtbaren Speckle-Muster die automatisierte Lokalisierung der Eckpunkte beeinträchtigen, muss der IR-Projektor während der Aufnahme abgedeckt und das Kalibriermuster stattdessen mit einer externen Infrarotlichtlampe bestrahlt werden (vgl. auch Abb. 3.7d). Wie

[2] http://www.vision.caltech.edu/bouguetj/calib_doc, letzter Zugriff: 14.02.2014

Tabelle 3.1.: Geschätzte Kameramatrizen und Verzeichnungsparameter für die aufgenommenen Farb- und Infrarotbilder.

	Farbbild	**Infrarotbild**
$\mathbf{K}$	$\begin{pmatrix} 523{,}73 & 0 & 308{,}66 \\ 0 & 520{,}86 & 266{,}90 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 584{,}70 & 0 & 328{,}53 \\ 0 & 582{,}00 & 242{,}89 \\ 0 & 0 & 1 \end{pmatrix}$
k_1, k_2, k_3	$\begin{pmatrix} 0{,}19873 & -0{,}41419 & 0 \end{pmatrix}$	$\begin{pmatrix} -0{,}09420 & 0{,}21353 & 0 \end{pmatrix}$
p_1, p_2	$\begin{pmatrix} -0{,}00240 & -0{,}00412 \end{pmatrix}$	$\begin{pmatrix} 0{,}00003 & 0{,}00044 \end{pmatrix}$

bereits erwähnt, entsprechen die von der Kinect gestreamten Bilder mit Bildmaßen von 640×480 Pixeln nicht der Auflösung der Kamerasensoren von 1280×1024 Pixeln. Es ist zu beachten, dass sich die aus der Kalibrierung ergebenden Parameter auf die reduzierten Bilder beziehen und nicht auf die Sensoren.

Zur Kalibrierung der Farb- und Infrarotkamera wurden 14 bzw. 25 Aufnahmen des Kalibriermusters aus verschiedenen Perspektiven gesammelt (siehe Abb. 3.10b). In der Tabelle 3.1 sind die im Rahmen der Kalibrierung geschätzten intrinsischen Parameter für das RGB- und das Infrarotbild aufgeführt. In beiden Fällen ist der Wert des Parameters k_3 auf Null gesetzt, da er nur für Objektive mit sehr starker Verzeichnung Relevanz hat (z.B. Fischaugenobjektiv). Ein geschätztes Modell der radialen und tangentialen Verzeichnung für die RGB-Bilder ist in der Abbildung 3.10c visualisiert. Beispielaufnahmen vor und nach der Entfernung der Bildverzeichnung sind in der Abbildung 3.11 gezeigt. Die aus der Kalibriermatrix abgeleitete Verschiebung $\Delta\mathbf{p}_0$ des Hauptpunktes entspricht beim Farbbild $\Delta x = -12$ und $\Delta y = 26$ Pixel und beim Infrarotbild $\Delta x = 8$ und $\Delta y = 2$ Pixel und ist vergleichbar mit den in der Literatur aufgeführten Zahlen für andere Kinectsysteme (siehe Tabelle B.2 im Anhang).

In der Tabelle 3.2 sind die extrinsischen Parameter aufgeführt. Sie beschreiben die Rotation und Translation zwischen der RGB- und der Tiefenkamera. Die Rotationsmatrix ähnelt einer Einheitsmatrix, was auf eine geringe Rotation zwischen beiden Kameras hinweist. Die Werte für die Translation in y- und z-Richtung liegen im ein- beziehungsweise zweistelligen Mikrometerbereich, während die ermittelte Verschiebung in x-Richtung bei 2,48cm liegt. Dieser Wert deckt sich mit dem an der Außenhülle der Kinect abschätzbaren Abstand zwischen den Kameras.

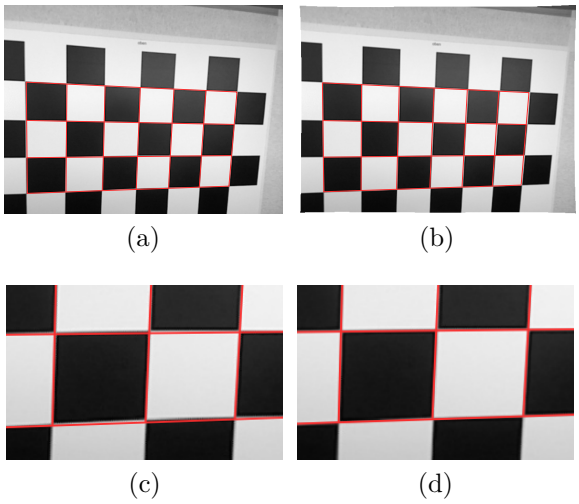


Abbildung 3.11.: **(a)** Originalaufnahme der RGB-Kamera. Zur Veranschaulichung der Verzerrungen wurden nachträglich gerade rote Linien eingezeichnet. **(b)** Entzerrte Aufnahme. **(c)**-**(d)** Detailausschnitte des mittigen, unteren Quadratpaares von (a) und (b) verdeutlichen die unterschiedliche Beugung der Quadratkanten. Im entzerrten Bild (d) gleichen sich diese an die gerade rote Linie an.

Tabelle 3.2.: Geschätzte extrinsische Parameter. Die Einträge des Translationsvektors sind in der Einheit Meter angegeben.

Rotationsmatrix R	**Translationsvektor t**
$\begin{pmatrix} 0{,}9999 & 0{,}0062 & 0{,}0079 \\ -0{,}0062 & 1{,}0000 & 0{,}0050 \\ -0{,}0079 & -0{,}0051 & 1{,}0000 \end{pmatrix}$	$\begin{pmatrix} -0{,}02481 & 0{,}00026 & 0{,}00003 \end{pmatrix}^T$

3.3.2. Bildtransformation und Punktwolkenerzeugung

Um die Aussagekraft der von der Kinect gestreamten Farb- und Tiefendaten zu erhöhen, werden diese kombiniert und in eine zusätzliche Darstellungsform überführt, der sogenannten colorierten Punktwolke. Dabei wird das 2.5D-Bild so transformiert, dass es deckungsgleich zum Farbbild ist und die Zuordnung korrespondierender Farb- und Tiefenpixel möglich wird. Beide Schritte stehen im Fokus dieses Unterabschnitts.

Dem Namen entsprechend bestehen Punktwolken aus Ansammlungen von Punkten, die aus der in Pixeleinheiten unterteilten zweidimensionalen Ebene herausgelöst sind und in einem dreidimensionalen, metrischen Koordinatensystem dargestellt werden. Die Achsen des Koordinatensystems sind in der Einheit Meter unterteilt. Aus jedem Pixel $\mathbf{p}_L = (u, v)^T$ ergibt sich ein dreidimensionaler Punkt $\mathbf{p}_{WL} = (X_L, Y_L, Z_L)^T$. Vorbereitend dazu werden die normalisierten Disparitäten des von der Kinect gestreamten

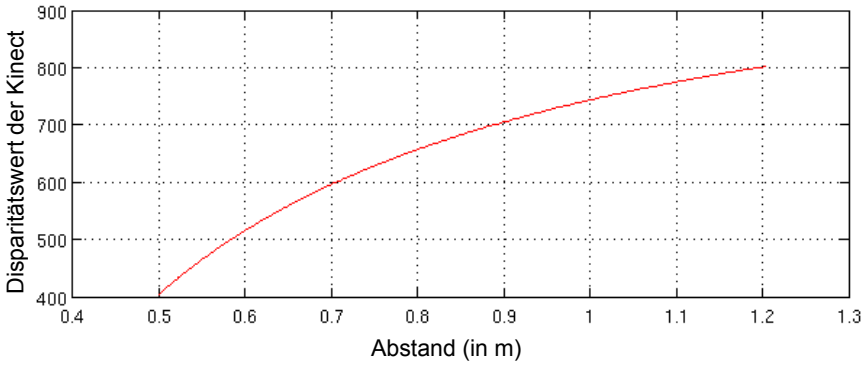


Abbildung 3.12.: Distanzwerte (in Meter) mit ihren korrespondierenden Disparitätswerten.

Tiefenbildes $\mathbf{I}_L$ in Metereinheiten umgewandelt. Daraus resultiert das Tiefenbild $\mathbf{I}_{mL}$. Die Berechnung folgt einer Näherung von Stéphane Magenat[3]:

$$\mathbf{I}_{mL}(u,v) = 0{,}1236 \cdot \tan\left(\frac{\mathbf{I}_L(u,v)}{2842{,}5} + 1{,}1863\right). \tag{3.14}$$

Der zulässige Messbereich der Kinect liegt zwischen 0,5 und 5 Metern und ist nicht linear auf den Wertebereich der Disparitäten verteilt. Dieser umfasst Integerwerte zwischen 0 und 2047. Werte ab 1022 repräsentieren Kamera-Objekt-Distanzen, die größer als 5 m sind. Werte ab 1093 ergeben laut Näherungsformel negative Werte. Der für diese Arbeit relevante Distanzbereich zwischen 0,5 und 1,2 Metern liegt im Disparitätswertebereich von 404 bis 802. Die Abbildung 3.12 zeigt den nichtlinearen Zusammenhang der beiden Datenrepräsentationen für den entsprechenden Wertebereich.

Im nächsten Schritt wird die Verzeichnung des Tiefenbildes unter Verwendung der Gleichungen 3.6 bis 3.9 herausgerechnet. Die entzerrten Koordinaten werden danach nicht in Pixelkoordinaten zurück transformiert, sondern bleiben in Form der verallgemeinerten homogenen Bildkoordinaten. Durch pixelweise Multiplikation mit dem Inhalt des Tiefenbildes $\mathbf{I}_{mL}$ werden die Koordinaten der Bildpunkte aus der zweidimensionalen projektiven Ebene in ein dreidimensionales System überführt:

$$\mathbf{p}_{WL} = \begin{pmatrix} X_L \\ Y_L \\ Z_L \end{pmatrix} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \cdot \mathbf{I}_{mL}(u,v) \tag{3.15}$$

Um jedem Punkt $\mathbf{p}_{WL}$ dieser Punktwolke einen Farbwert zuordnen zu können, werden RGB- und Tiefenbild miteinander registriert. Dazu wird der Punkt $\mathbf{p}_{WL}$, entspre-

[3]http://openkinect.org/wiki/Imaging_Information, letzter Zugriff: 24.07.2014

chend der Gleichung 3.10 und unter Verwendung der extrinsischen Parameter aus der Tabelle 3.2, vom Koordinatensystem der linken Tiefenkamera in das Koordinatensystem der rechten Farbkamera überführt:

$$\mathbf{p}_{WR} = \begin{pmatrix} X_R \\ Y_R \\ Z_R \end{pmatrix} = \mathbf{R} \cdot \begin{pmatrix} X_L \\ Y_L \\ Z_L \end{pmatrix} + \mathbf{t}. \tag{3.16}$$

Im nächsten Schritt wird der Punkt $\mathbf{p}_{WR}$ auf den Bildsensor der Farbkamera projiziert, wobei $\mathbf{K}_{RGB}$ der Kameramatrix der RGB-Kamera entspricht:

$$\begin{pmatrix} \lambda u' \\ \lambda v' \\ \lambda \end{pmatrix} = \mathbf{K}_{RGB} \cdot \mathbf{p}_{WR}. \tag{3.17}$$

Eine nachgeschaltete Division mit λ ergibt die entsprechenden Koordinaten (u', v') des Farbbildes $\mathbf{I}_R$. Auf diese Weise kann dem dreidimensionalen Punkt $\mathbf{p}_{WR}$ der passende Farbwert $\mathbf{I}_R(u', v')$ zugeordnet werden.

Zugleich kann die gewonnene Information auch dazu eingesetzt werden um das Tiefenbild $\mathbf{I}_{mL}$ so zu transformieren, dass es deckungsgleich zum Farbbild ist. Das neue Tiefenbild $\mathbf{I'}_{mL}$ ergibt sich nun durch folgende Zuweisung:

$$\mathbf{I'}_{mL}(u', v') = \mathbf{I}_{mL}(u, v). \tag{3.18}$$

Wie bereits erwähnt, wird in dieser Arbeit, analog zu [Khoshelham und Elberink, 2012], das Infrarotbild zur Kalibrierung verwendet, weil das Tiefenbild keine Eckenlokalisierung innerhalb des planaren Schachbrettmusters erlaubt. Diese Vorgehensweise benötigt eine vorgeschaltete Anpassung der Bilddaten, da $\mathbf{I'}_{mL}$ und $\mathbf{I}_R$ trotz Kalibrierung einen Shift aufweisen. Dabei ist das Tiefenbild um $\delta x = -3$ und $\delta y = -3$ Pixel nach links oben verschoben. Ein Vergleich des Epipolarlinienverlaufs im rektifizierten Farb- und Infrarotbild liefert jedoch keinen Hinweis auf eine missglückte Kalibrierung (siehe Abbildung 3.13). Als weiterer Grund kommt eine Verschiebung zwischen Infrarot- und Tiefenbild in Frage. Diese Erklärung deckt sich auch mit den Erkenntnissen von [Smisek et al., 2013] und [Khoshelham und Elberink, 2012], wobei letztere lediglich eine horizontale Verschiebung ausgemacht haben. Um für die nachfolgenden Schritte eine vollständige Deckungsgleichheit zu gewährleisten, wird das 2.5D-Bild vor allen Transformationsoperationen und vor Entfernung der Bildverzeichnung um δx und δy verschoben, damit es mit dem Infrarotbild übereinstimmt.

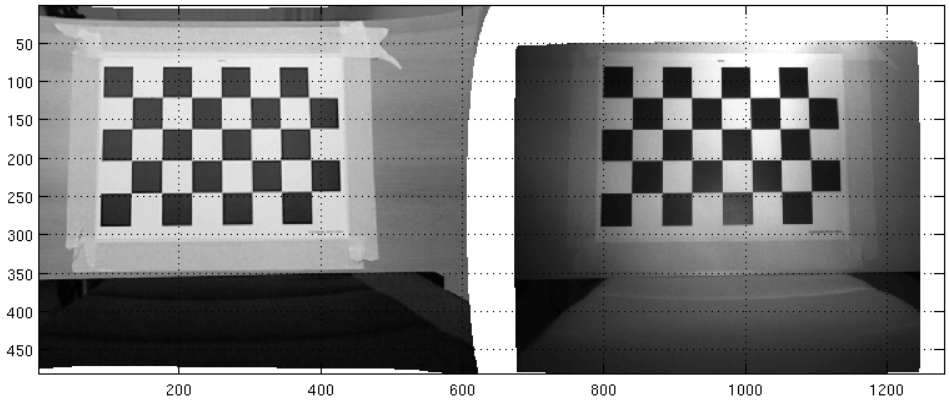


Abbildung 3.13.: Rektifiziertes Farb- (l.) und Infrarotbild (r.). Die gestrichelten horizontalen Linien visualisieren einzelne begradigte Epipolarlinienverläufe. Diese stimmen zwischen den beiden Bildern überein (siehe z.B. $y = 350$), weshalb eine missglückte Kalibrierung als Grund für die Verschiebung zwischen Farb- und Tiefenbild ausgeschlossen wird.

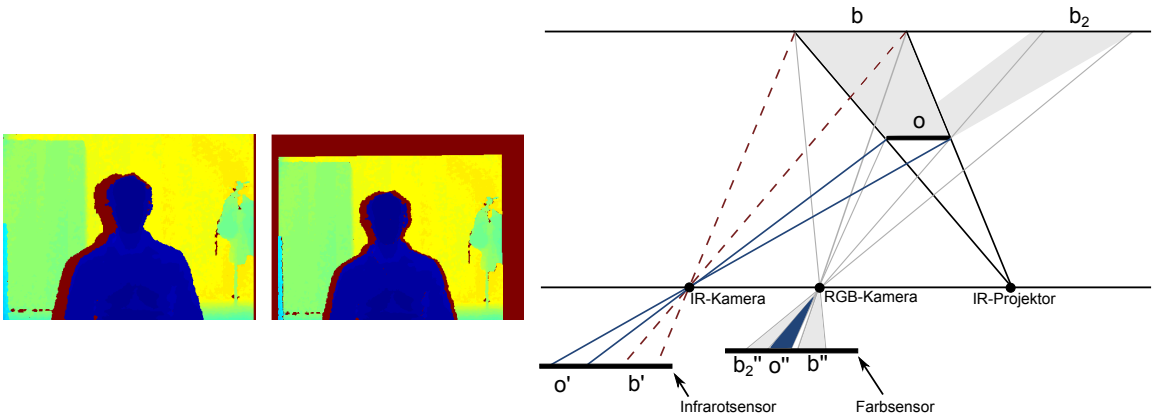
3.4. Entfernung der Tiefenwertfehler

Die 2.5D-Tiefenbilder enthalten Pixel, für welche keine Distanzwerte berechnet werden konnten. Diese werden in der Literatur als *nmd-Pixel* bezeichnet (*engl.* no measured depth) ([Yu et al., 2013], [Camplani und Salgado, 2012]). Ihre Enstehung und Korrektur steht im Zentrum dieses Unterkapitels.

3.4.1. Entstehung der nmd-Pixel

Die Entstehung der nmd-Pixel hat unterschiedliche Gründe, die mit darüber entscheiden, ob die Fehlerwerte korrigiert oder im Bild belassen werden. In den Abbildungen 3.14a und 3.14b sind zwei Tiefenbilder gezeigt, bei denen die nmd-Pixel dunkelrot dargestellt sind. Die erste Abbildung entspricht der Ursprungsversion, welche aus dem Speckle-Muster berechnet wird. Die zweite Abbildung ist dessen Transformation und Projektion auf den Sensor der Farbkamera und somit deckungsgleich zum RGB-Bild. Auffallend ist der dunkelrote Rahmen auf der linken, oberen und rechten Bildseite. Dieser entsteht durch *fehlende Szeneninformation* bei der Projektion der Bilddaten zwischen den beiden Kamerasensoren. So nimmt der Tiefensensor aufgrund seiner größeren Brennweite im Vergleich zum RGB-Sensor Szeneninformation in größerem Maßstab jedoch mit kleinerem Bildausschnitt auf. Eine Extrapolation der fehlenden Tiefeninformation ist nicht erforderlich, da die Merkmalsextraktion auf das Personenareal beschränkt ist.

Ein weiterer Grund für das Auftreten von nmd-Pixeln ist die *Nichteinhaltung des*



(a) (b) (c)

Abbildung 3.14.: **(a)** Von der Kinect ausgegebenes 2.5D-Bild (nicht deckungsgleich zum RGB-Bild). **(b)** Auf den Farbsensor projizierte Tiefeninformation. **(c)** Die Grafik verdeutlicht, weshalb der Schatten im Ursprungstiefenbild auf der linken und im projizierten auf beiden Seiten auftritt.

erlaubten Aufnahmeabstandes, welcher im Bereich von 0,5 bis 5 m liegt [Yu et al., 2013]. Dieser Fall ist für das vorgesehene Szenario nicht von Bedeutung, da die Platzierung des Patienten innerhalb eines Distanzintervalls von 0,5 bis 1.2 m zur Kamera vorgesehen ist.

Ebenfalls zu Fehlerwerten führen *materialtypische Eigenschaften*, zum Beispiel bei spiegelnden Flächen oder transparenten Materialien wie Glas ([Yu et al., 2013], [Danciu et al., 2012]). Auch diese spielen eine eher geringe Rolle und werden im Folgenden nicht gesondert betrachtet.

Die in dieser Arbeit vorrangige Ursache von nmd-Pixeln ist die *Verdeckung* von Objekten durch andere Objekte. Die dadurch entstehenden Areale von ungültigen Werten werden als Schatten bezeichnet. Die Abbildung 3.14c visualisiert die Entstehung dieser Schatten und verdeutlicht, warum der Schatten im originalen 2.5D-Bild auf einer Seite und im transformierten 2.5D-Bild auf beiden Seiten der Person auftritt. Der von der Person aus gesehen einseitige Schatten entsteht bereits im Zuge der Datenaufnahme. Ein Objekt o wird vom Infrarotprojektor mit einem Speckle-Muster bestrahlt. Weil das Objekt einen Teil des Hintergrundes verdeckt, treffen auf diesen keine Speckle-Punkte auf und ein Schatten b entsteht. Da die Infrarotkamera im Vergleich zum Projektor räumlich versetzt ist, erfasst ihr Aufnahmefeld diesen Schattenbereich. Aufgrund des fehlenden Speckle-Musters können für diesen jedoch keine Tiefenwerte berechnet werden. Neben einem Abbild o' des Objekts, trifft auch ein Abbild b' des Schattens auf den Infrarotsensor. Der Bereich b_2 wird vom Objekt

o verdeckt und infolgedessen nicht auf den Sensor abgebildet. Das 2.5D-Bild wird im nächsten Schritt zur Registrierung der Farb- und Tiefeninformation weiterverwendet, indem seine Pixelkoordinaten in Weltkoordinaten transformiert und ihre Tiefenwerte auf die korrespondierenden Pixelkoordinaten des räumlich versetzten Farbsensors projiziert werden. Die Hintergrundareale b_2 und b werden als nmd-Regionen auf die Bildebene abgebildet, wodurch links und rechts vom Objektabbild o'' die Schatten b'' und b_2'' entstehen. Diese Schatten werden im Folgenden korrigiert, da sie direkt an die Person angrenzen und die nachgeschaltete Datenanalyse beeinträchtigen können.

3.4.2. Literaturauswertung

In der Literatur finden sich verschiedene Möglichkeiten der nmd-Pixel Entfernung. [Lai et al., 2011] verwenden zur Interpolation der ungültigen Tiefenwerte einen rekursiven Medianfilter der Größe 5×5. Es werden nur die gültigen Pixelwerte in die Berechnung einbezogen und der Filter so lange über das Bild geschoben, bis alle fehlenden Pixelwerte gefüllt sind. Auf ihre Vorgehensweise in Übergangsbereichen, beispielsweise um zu vermeiden, dass Kanten zwischen Vordergrund und Hintergrund verwischt werden, gehen sie nicht näher ein.

[Matyunin et al., 2011] beziehen neben dem Tiefenbild auch das Farb- beziehungsweise Grauwertbild in die Datenrekonstruktion ein, weil sie die Annahme zu Grunde legen, dass zusammenhängende Regionen ähnlicher Farbe einen ähnlichen Tiefenwert besitzen. Der fehlende Wert wird aus dem Median der Nachbartiefenwerte ermittelt, wenn die Anzahl der Nachbarpixel mit gültigem Tiefenwert einen bestimmten Grenzwert nicht unterschreitet und die Intensitätsdifferenz zu den Nachbarpixeln im Farb- und Grauwertbild einen bestimmten Wert nicht überschreitet. Genaue Zahlen für die Grenzwerte werden jedoch nicht angegeben.

[Camplani und Salgado, 2012] verwenden rekursiv Informationen aus mehreren zeitlich aufeinanderfolgenden Tiefenbildern, um die fehlenden Tiefenwerte zu rekonstruieren. Die Häufigkeit des Auftretens eines bestimmten Tiefenwertes dient dabei als Maß für seine Verlässlichkeit. Der Vorgang der Datenrekonstruktion wird mit einer Datenglättung kombiniert, welche die Ähnlichkeit der Nachbarpixel im Farb- und Tiefenbild einbezieht. Diese Vorgehensweise bezeichnet man als rekursives Joint-Bilateral-Filtering.

[Danciu et al., 2012] stellen einen Suchalgorithmus vor, der das Tiefenbild reihenweise durchläuft und für jede Reihe sogenannte Zero-Crossings bestimmt. Dabei handelt es sich um die Positionen ungültiger Pixelwerte, an welche Pixel mit gültigen Tiefenwerten angrenzen. Basierend auf der Annahme, dass sich der Schatten auf dem

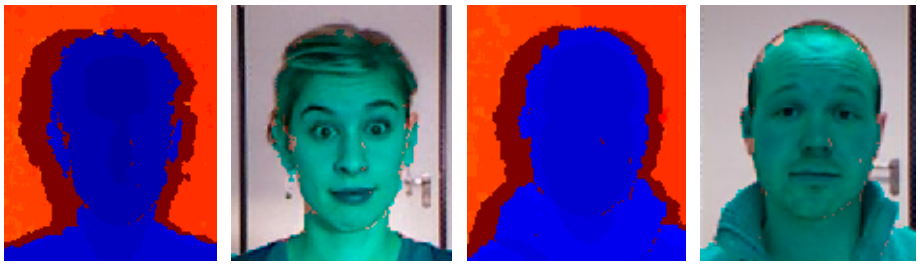
weiter entfernten Objekt, also dem Hintergrund, befindet, wird den Zero-Crossing-Pixeln der Tiefenwert der angrenzenden Region mit der größten Distanz zugeordnet.

[Yu et al., 2013] stellen einen Ansatz zur Detektion der Schattenregionen vor, in Abgrenzung zu beispielsweise materialbedingten nmd-Pixeln, und füllen diese ebenfalls mit den Werten des Hintergrundes. Auch sie gehen reihenweise vor und berechnen den Wert eines nmd-Pixels aus seinen sechs horizontal benachbarten, gültigen Hintergrundpixeln. Um den Einfluss des Rauschens in diesen Referenzwerten zu verringern, entfernen sie den maximalen und minimalen Tiefenwert und verwenden den Mittelwert der vier verbleibenden Nachbarpixel, um für das nmd-Pixel einen Tiefenwert zu schätzen.

3.4.3. Eigener Ansatz

Die in dieser Arbeit auftretenden nmd-Pixel sind nicht ausschließlich dem Hintergrund zuzuordnen, wie die Abbildungen 3.15a bis 3.15d verdeutlichen. Die Ursachen sind zum einen Verdeckungen innerhalb des Gesichts und zum anderen materialtypische Eigenschaften. Eine Vorgehensweise nach [Danciu et al., 2012] und [Yu et al., 2013] würde jedoch auch diesen im Konturbereich des Gesichtsareals liegenden nmd-Pixeln Tiefenwerte des Hintergrunds zuordnen. Aus diesem Grund wurde im Rahmen dieser Arbeit eine Methode untersucht, bei der, basierend auf den Farbwerten der nmd-Pixel und den Tiefenwerten der Nachbarpixel, die Zugehörigkeit der einzelnen Pixel zu Vordergrund oder Hintergrund bestimmt wird. Es zeigte sich allerdings, dass durch ähnliche Vordergrund- und Hintergrundfarben, wie z.B. einem grauen Pullover vor grauem Hintergrund, Fehlzuordnungen entstehen können und der insgesamte Nutzen in geringem Verhältnis zu Rechendauer und Aufwand steht. Da die entsprechenden Pixel auch häufig in weniger relevanten Randbereichen liegen, wurde diese Vorgehensweise wieder verworfen.

Das letztendlich konzipierte und umgesetzte Verfahren ist daher unabhängig vom Farbwert und stützt sich ausschließlich auf die Nachbarschaftsbeziehungen. Der Ablauf lässt sich in zwei Hauptschritte unterteilen. Im ersten Schritt wird die *Zuordnung* jedes nmd-Pixels zu einer von zwei möglichen Regionen ermittelt (Personen- oder Hintergrundregion). Im zweiten Schritt wird der fehlende Tiefenwert dieses Pixels *interpoliert*, indem die Tiefenwerte der gültigen Nachbarpixel mit derselben Regionszuordnung gemittelt werden, wobei eine bestimmte Mindestanzahl an geeigneten Nachbarpixeln gegeben sein muss. Die Abbildung 3.16 beinhaltet eine Übersicht der notwendigen Schritte, auf die im Folgenden näher eingegangen wird.



(a) (b) (c) (d)

Abbildung 3.15.: **(a)-(d)** Korrespondierende 2.5D- und RGB-Bilder. Die mittels Vordergrund-Hintergrund-Segmentierung eingegrenzten Personenareale sind blau, die nmd-Areale rotbraun dargestellt. Eine blaue Markierung dieser Personenareale im RGB-Bild verdeutlicht, dass nmd-Areale auch innerhalb des Gesichts auftreten.

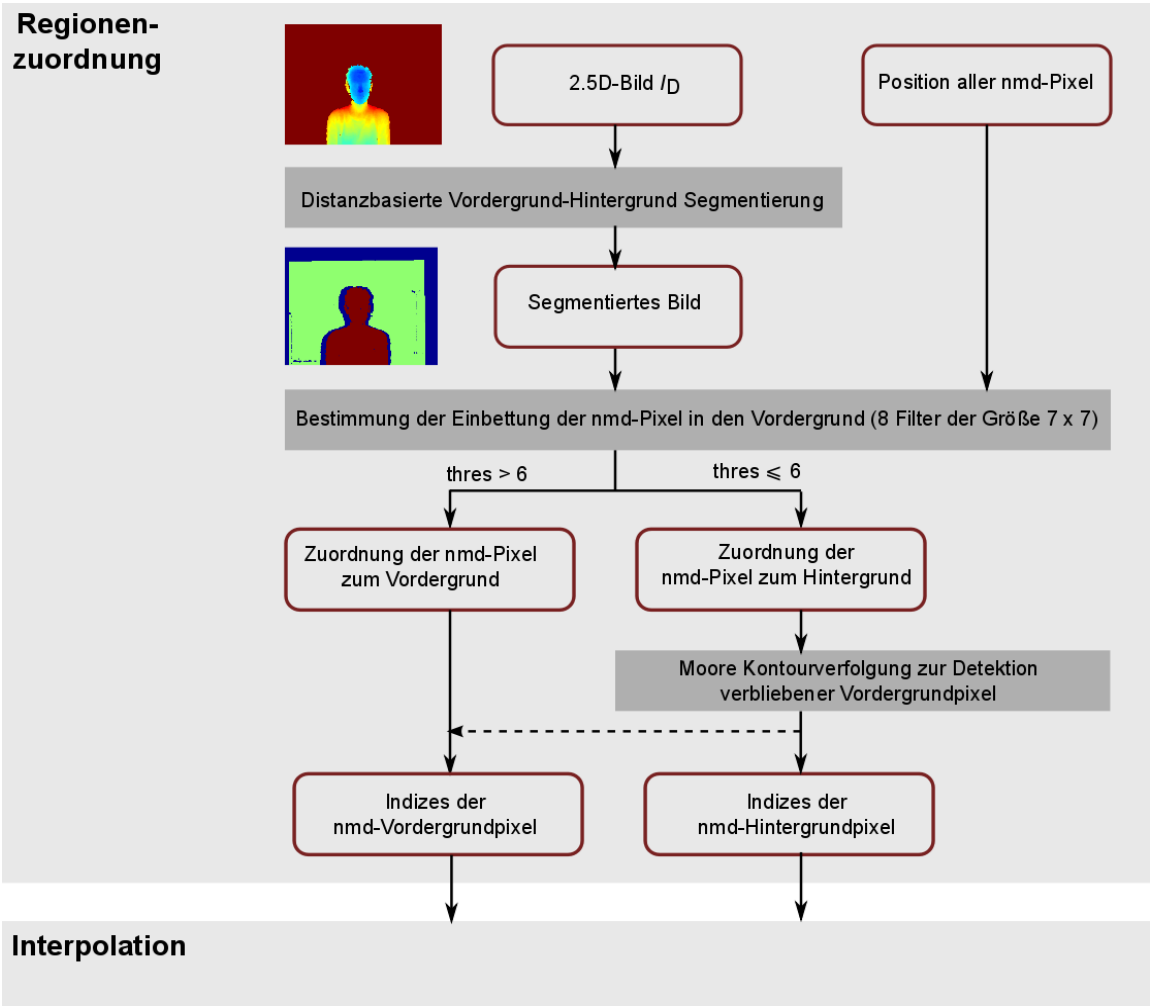
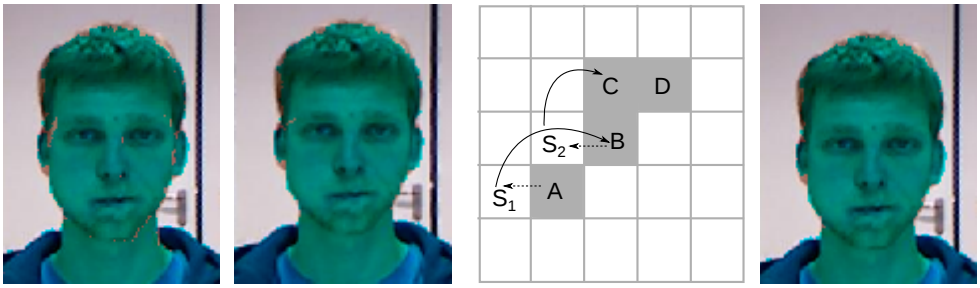


Abbildung 3.16.: Übersicht über die Schritte der nmd-Pixel-Regionenzuordnung. Rechtecke mit abgerundeten Ecken symbolisieren Daten, Rechtecke mit spitzen Ecken Arbeitsschritte.



(a) (b) (c) (d)

Abbildung 3.17.: **(a)** Ausgangsbild mit vollumschlossenen und halbinselförmigen nmd-Regionen. **(b)** Personenregion nach Bestimmung des Einbettungsgrades. Komplett umschlossene nmd-Pixel wurden erfolgreich der Personenregion zugeordnet und Halbinseln so verschlossen, dass sie im nächsten Schritt der Personenregion zugeordnet werden können. **(c)** Visualisierung des Moore-Konturverfolgungsalgorithmus. **(d)** Resultierende Personenregion nach Zuordnung der nmd-Pixel.

Zuordnung

Für die Zuordnung sind, neben dem 2.5D-Bild, die Positionen der nmd-Pixel im Bild erforderlich. Im ersten Schritt werden alle Pixel mit gültigem Tiefenwert in Personen- und Hintergrundregion segmentiert. Diese Segmentierung erfolgt distanzbasiert und stützt sich auf die vorgegebene Positionierung der übenden Person in einem Distanzintervall von 0,5 bis 1,2 Metern. Bei der Zuordnung werden zwei Fälle unterschieden. Nmd-Pixel, die vollständig von einer der beiden Regionen umschlossen sind und somit wie eine Insel darin eingebettet liegen, werden automatisch der sie umgebenden Region zugeordnet. Pixel, die im Grenzbereich zwischen Personen- und Hintergrundregion liegen, werden vorrangig dem Hintergrund zugeordnet, da diese ungültigen Werte vorwiegend durch die Verdeckung des Hintergrundes zustande kommen. Die Abbildung 3.17a verdeutlicht jedoch, dass diese, auch von [Danciu et al., 2012] und [Yu et al., 2013] getroffene, Annahme nicht in allen Fällen zutrifft. Die halbinselförmig in die Personenregion ragenden nmd-Regionen entstehen insbesondere durch Rauschen an Objektgrenzen und durch Verdeckungen innerhalb eines (Vordergrund-)Objektes. Eine pauschale Zuordnung dieser im Konturbereich liegenden Pixel zum Hintergrund wäre daher fehlerhaft, weshalb für diese Pixel der *Grad der Einbettung* in den Vordergrund bestimmt bestimmt wird. Dieser bildet die Entscheidungsgrundlage für die Zuordnung zum Vorder- oder Hintergrund.

Der Grad der Einbettung wird auf folgende Art und Weise bestimmt. Im Anschluss an die Binarisierung des 2.5D-Bildes, bei welcher allen Vordergrundpixeln der Wert 1 und den restlichen nmd- und Hintergrundpixeln den Wert 0 zugeordnet wird, wird

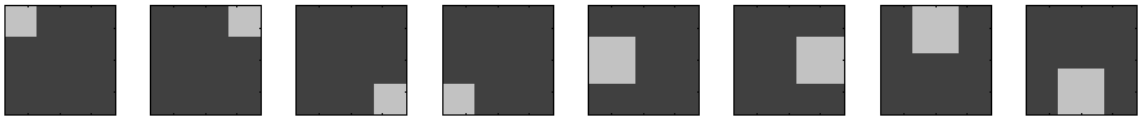


Abbildung 3.18.: Alle acht binären 7×7 Filter, die zur Bestimmung der Stärke der Einbettung in den Vordergrund eingesetzt werden. Hellgraue Felder entsprechen einem Wert von 1, die dunkelgrauen einem Wert von 0. Die kleinen hellgrauen Felder haben eine Größe von 2×2, die großen von 3×3 Pixeln.

jedes nmd-Pixel mit acht verschiedenen Filtern der Größe $n \times n$, mit $n = 7$, gefiltert. Die Abbildung 3.18 zeigt die verwendeten Filter, die aus den Werten 0 (dunkelgrau) und 1 (hellgrau) bestehen. Jedes Filter gibt dabei nur eine binäre Antwort aus. Sie ist 0, wenn die Multiplikation der Filterelemente mit den korrespondierenden Bildpixeln in der Summe den Wert 0 ergibt und 1, sobald die Summe größer beziehungsweise gleich 1 ist.

Die einzelnen Antworten der acht Filter werden aufsummiert, sodass als Ergebnis ein Wert s zwischen 0 oder 8 ausgegeben wird. Dieser dient als Entscheidungswert. Bei einem Wert $s > 6$ wird das nmd-Pixel der Personenregion zugeordnet, bei $s \leq 6$ dem Hintergrund. Diese Vorgehensweise führt dazu, dass eng vom Vordergrund umschlossene nmd-Pixel identifiziert und diesem zugeordnet werden können. Kleinere, in die Personenregion ragende nmd-Halbinseln werden so entfernt oder zumindest an schmalen Trennstellen, zum Beispiel zwischen Kopf und Ohr, verbunden. Dadurch entstehen geschlossene Inseln, die dann eindeutig der Personenregion zugeordnet werden können. Aufgrund der beschränkten Filtergröße von $n = 7$ eignet sich die Methode jedoch nicht zur Bestimmung der Einbettung von Pixeln, die im Zentrum größer Inselbereiche liegen, wie Abbildung 3.17b zeigt. Sie werden in einem nachgeschalteten Schritt gesondert lokalisiert. Dies geschieht mit Hilfe des *Moore-Konturverfolgungsalgorithmus* (*engl.* Moore Contour Tracing) [Reddy et al., 2012]. Dabei wird Pixel für Pixel die Kontur eines Objektes, in diesem Fall der nmd-Pixel-Ansammlung, verfolgt. Die Abbildung 3.17c visualisiert die Vorgehensweise. Die grauen Pixel repräsentieren das Objekt, die weißen den Hintergrund. Die Startposition S_1 wird so initialisiert, dass sie direkt an ein Konturpixel angrenzt (z.B. A), jedoch nicht auf einem Pixel des Objekts liegt. Das entsprechende Konturpixel wird nun im Uhrzeigersinn umschritten. Sobald ein weiterer Konturpunkt (hier B) angetroffen wird, wird dieser ebenfalls im Uhrzeigersinn umlaufen, wobei der neue Startpunkt S_2 das zuletzt besuchte weiße Pixel ist. Diese Vorgehensweise wird so lange wiederholt, bis das *Jacob's Stopping Kriterium* erfüllt ist [Reddy et al., 2012]. Dieses beschreibt den Fall, dass ein Objektpixel ein

zweites Mal von derselben Richtung aus beschritten wird.

Die durch die Konturverfolgung ermittelten Inseln werden abschließend der sie umgebenden Region zugeordnet. Die Abbildung 3.17d visualisiert das Endergebnis des Zuordnungsverfahrens anhand eines Beispielbildes. Nachdem die Zuordnung abgeschlossen ist, kann mit der Interpolation der nmd-Pixel begonnen werden.

Interpolation

Der Tiefenwert eines nmd-Pixels wird interpoliert, indem ihm der arithmetische Mittelwert aus den Tiefenwerten seiner $w \times w$ -Nachbarschaft zugewiesen wird. Bei dieser Zuweisung gelten jedoch zwei Einschränkungen. Zum einen werden ausschließlich die Nachbarpixel in die Berechung einbezogen, die derselben Region angehören (Personen- oder Hintergrundregion). Zum anderen muss eine Mindestanzahl geeigneter Nachbarpixel vorhanden sein. Dies soll sicherstellen, dass die Interpolationsergebnisse auf einer ausreichenden Datenmenge basieren und nicht durch einzelne Rauschpixel beeinträchtigt werden.

Entsprechend dieser Vorgehensweise werden alle nmd-Pixel zum Zwecke der Interpolation in einer Schleife durchlaufen. Aufgrund der Einschränkungen kann es jedoch vorkommen, dass nach einem Durchlauf nmd-Pixel verbleiben, deren Nachbarschaftspixel die genannten Kriterien nicht erfüllen. In diesem Fall wird ein weiterer Durchlauf unter Einbeziehung der neu interpolierten Pixel nachgeschaltet. Dies geschieht so oft, bis die Menge der verbliebenen nmd-Pixel leer ist oder bei zwei aufeinanderfolgenden Durchläufen konstant bleibt. Letzteres tritt beispielsweise in Randbereichen auf, in denen das Interpolationsfenster in eine andere Region ragt. Dadurch kann die erforderliche Mindestanzahl an geeigneten Nachbarpixeln nicht erreicht werden. Ist dies der Fall, wird diese sukzessive reduziert. Die Beschränkung auf Pixel derselben Regionszuordnung bleibt bestehen.

3.5. Zusammenfassung

Der Fokus dieses Kapitels lag auf der Aufnahme, Vorverarbeitung und Korrektur der Tiefendaten. Die Tiefendaten bilden die zentrale Ausgangsbasis der Feedbackerzeugung. Die Vorgehensweise setzte sich weitestgehend aus der Kombination bereits existierender Verfahren zusammen. Die Intention dieses Kapitels war somit vorwiegend die Dokumentation der durchgeführten Arbeitsschritte und Ergebnisse. Auf diese Weise soll, in Kombination mit den anderen Kapiteln dieser Arbeit, eine vollständige Übersicht über alle Teilschritte des technischen Gesamtablaufs ermöglicht werden.

4. Merkmalsextraktion

Um die Übungsdurchführung eines Patienten zu bewerten, soll diese mit den Referenzdaten von gesunden Personen verglichen werden. Zu diesem Zweck ist die Wahl geeigneter Merkmale von zentraler Bedeutung. Die Einordnung der Merkmalsextraktion in den technischen Gesamtablauf ist in der Abbildung 4.1 zu sehen.

Die einleitende Literaturübersicht dient als Ausgangsbasis für die Wahl von fünf verschiedenen Merkmalsextraktionsverfahren (Merkmalstypen). Letztere werden in den Unterkapiteln 4.3 bis 4.6 einzeln vorgestellt und experimentell evaluiert. Allgemeine Informationen zum Testszenario finden sich vorgeschaltet im Unterkapitel 4.2. Das abschließende Unterkapitel beinhaltet eine Zusammenfassung der Einzelevaluationsergebnisse, sowie einen darauf aufbauenden Leitfaden zur Merkmalsauswahl.

4.1. Literaturübersicht zu Tiefendatenmerkmalen

Auch wenn die Entwicklung von automatisierten, therapeutischen Mimiktrainern noch in den Kinderschuhen steckt, existieren bereits zahllose wissenschaftliche Publikationen, die sich mit der automatisierten Analyse von Gesichtern im Allgemeinen beschäftigen. Eine vollständige Auflistung dieser Ansätze übersteigt den Rahmen der vorliegenden Arbeit, zumal sich nicht alle gleichermaßen für die Integration in eine therapeutische Trainingsplattform eignen. Aus diesem Grund ist die Literaturübersicht in zwei Teile gegliedert. Im ersten Teil werden in Frage kommende Verfahren identifiziert und anschließend im Abschnitt 4.1.2 ausführlicher vorgestellt. Der letzte Abschnitt 4.1.3 fasst die Ergebnisse zusammen.

4.1.1. Eingrenzung und Identifikation geeigneter Verfahren

Die zentralen Anwendungsbereiche der maschinenbasierten Gesichtsanalyse lassen sich in drei Gruppen unterteilen. Dabei unterscheidet man zwischen der Gesichtserkennung (*engl.* face recognition), der Klassifikation von Gesichtsausdrücken (*engl.* facial expression recognition, *Abk.* FER) und der Erkennung von Basisbewegungen (Action-Units, *Abk.* AU) (siehe dazu auch Abschn. 5.1.2). Die Gesichtserkennung umfasst zwei

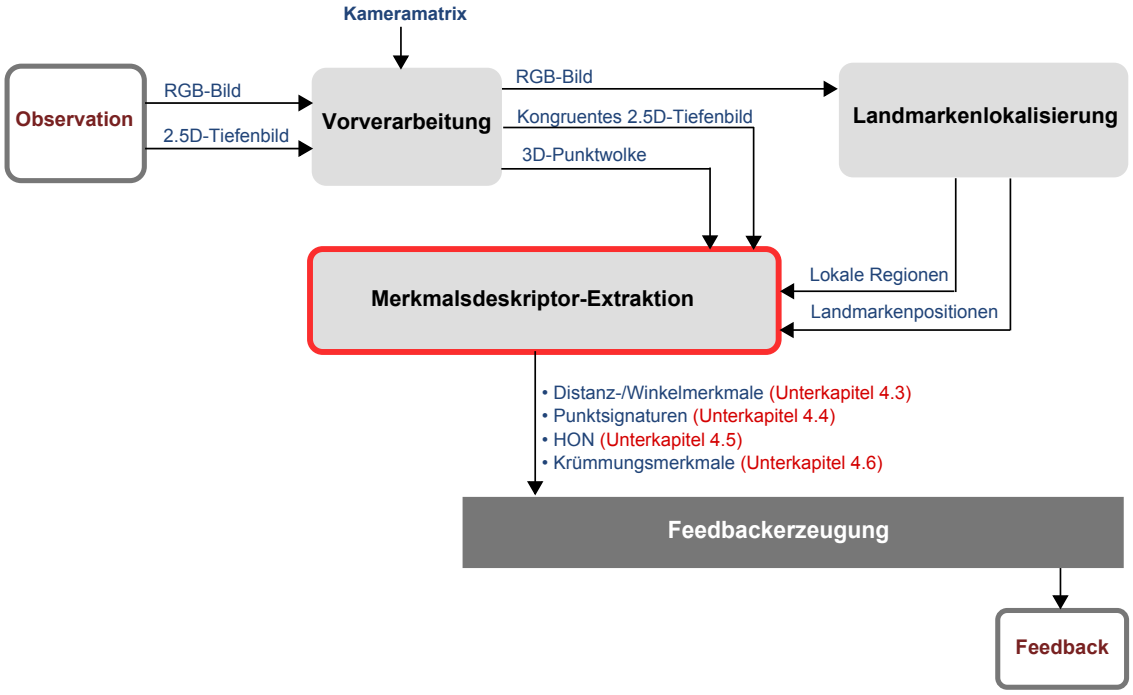


Abbildung 4.1.: Positionierung der Merkmalsextraktion innerhalb der technischen Gesamtarchitektur.

verschiedene Teilanwendungen, die unter den Begriffen Identitätsbestimmung und Verifizierung zusammengefasst werden[1].

Beide Zielstellungen profitieren in der Regel von einer Invarianz gegenüber Gesichtsausdrücken, um Personen in verschiedensten Situationen robust erkennen zu können. Oftmals konzentrieren sich entsprechende Verfahren auf die obere Gesichtshälfte, da diese weniger stark von Bewegungen der Mimik beeinflusst wird [Abate et al., 2007]. Da das Ziel der vorliegenden Arbeit aber gerade in der Erkennung und Bewertung von Gesichtsmimik besteht, liegt der Fokus der folgenden Literaturübersicht auf Veröffentlichungen zur Mimik- und AU-Erkennung.

Auf Basis der Literaturrecherche lassen sich die zu extrahierenden Merkmalsdeskriptoren in unterschiedliche Kategorien einteilen. Aufgrund der hohen Anzahl kann nur eine kleine Untermenge der Merkmale in die anschließende experimentelle Auswertung einbezogen werden. In Abbildung 4.2 ist eine entsprechende Systematik zu sehen, wobei die in dieser Arbeit gewählten Kategorien mit einer roten Umrandung markiert sind.

[1]Bei der Identitätsbestimmung werden die aus einem Probebild extrahierten Merkmale mit denen einer Datenbank verglichen, um einen korrespondierenden Datenbankeintrag und somit die wahrscheinlichste Identität der Person zu ermitteln. Bei der Verifizierung wird eine bestimmte Identität vermutet und über einen Merkmalsabgleich mit dem entsprechenden Datenbankeintrag bestätigt oder widerlegt.

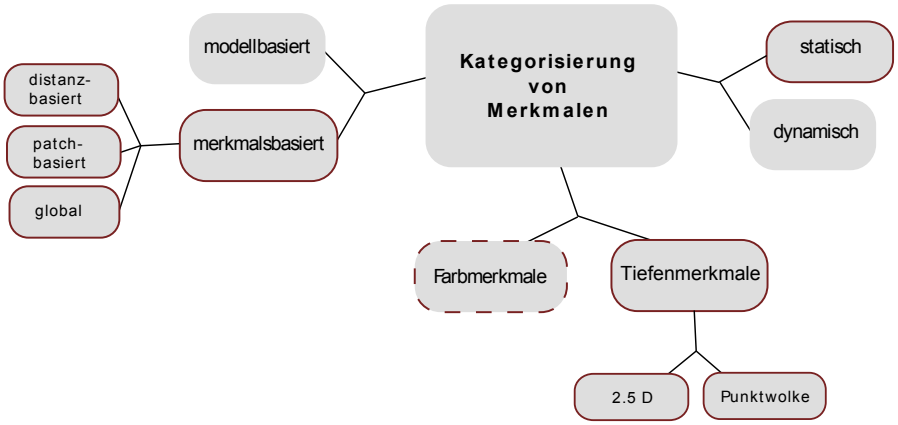


Abbildung 4.2.: Kategorisierung der verschiedenen Merkmalsarten. Diese Arbeit beschränkt sich auf den Einsatz von Merkmalen, die den rot umrandeten Kategorien angehören. Tiefenmerkmale stehen im Mittelpunkt, werden aber für die Aufgabenstellung der automatisierten Landmarkenlokalisierung um Farbinformationen ergänzt.

Die Unterscheidung zwischen *statischen* und *dynamischen* Merkmalen ist sowohl auf Farb- als auch Tiefenmerkmale anwendbar. Statische Merkmalsdeskriptoren werden aus Einzelbildern, dynamische aus Bildsequenzen extrahiert (u.a. [Sun und Yin, 2008]). Diese Arbeit beschränkt sich auf statische Merkmale, da die ausgewählten Fazialisübungen vornehmlich statischer Natur sind. Zusätzlich ist im vorgesehenen Szenario die Ausführgeschwindigkeit der Übungen weniger entscheidend als das konstante Halten über einen bestimmten Zeitraum. Eine spätere Erweiterung um eine Kombination der Einzelübungen zu einem dynamischen Ablauf ist möglich und sinnvoll, übersteigt jedoch den Rahmen dieser Arbeit.

Auf die verschiedenen Möglichkeiten der Datenrepräsentation, insbesondere in Form von *Farb-* oder *Tiefendaten*, wurde bereits im Kapitel 3 eingegangen. Der Schwerpunkt dieser Arbeit liegt auf der Tiefeninformation und ihren Varianten.

Abschließend wird zwischen *modellbasierten* (*engl.* model-based) und *merkmalsbasierten* (*engl.* feature-based) Extraktionsverfahren unterschieden [Fang et al., 2011]. Bei den modellbasierten Verfahren wird anhand von Trainingsbeispielen ein statistisches Modell gelernt, welches bestimmte Eigenschaften des Gesichts, wie etwa die Topologie und Verteilung der Landmarken, repräsentiert (u.a. [Ramanathan et al., 2006], [Mpiperis et al., 2008]). Die Anpassung des Modells an ein neues Gesicht, zum Beispiel durch Minimierung einer Distanzfunktion, beeinflusst die Modellparameter, welche dann als Merkmalsvektor an einen Klassifikator übergeben werden können. Da 3D-Modelle häufig viele Freiheitsgrade besitzen, ist die Anpassung vergleichsweise komplex und rechenintensiv. Beim merkmalsbasierten Ansatz entfällt der Zwischen-

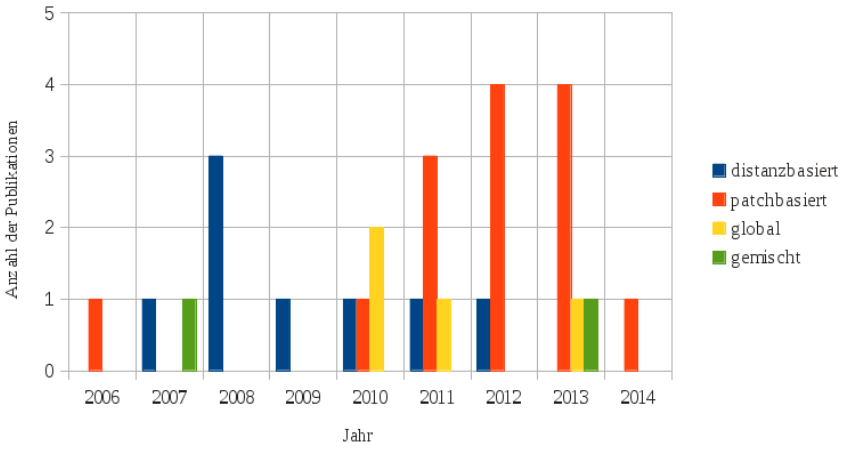


Abbildung 4.3.: Zuordnung der Publikationen zu ihrem jeweiligen Publikationsjahr. Von 28 Veröffentlichungen werden in acht distanzbasierte, in 14 patchbasierte und in vier globale Verfahren eingesetzt. Die verbliebenen zwei Veröffentlichungen sind den gemischten Verfahren zuzuordnen, da distanz- und patchbasierte Vorgehensweisen miteinander kombiniert werden.

schritt über ein Modell und die Informationen werden direkt aus den Eingangsdaten extrahiert. Dies ermöglicht eine gezielte Informationsextraktion aus einzelnen Regionen und darüber hinaus eine unkomplizierte Kombination mehrerer Verfahren. Da in dieser Arbeit gezieltes, regionenbezogenes Feedback angestrebt wird, konzentrieren sich die nachfolgenden Analysen auf merkmalsbasierte Ansätze.

4.1.2. Literaturübersicht

Die folgende Literaturübersicht stützt sich auf die Auswertung von 28 wissenschaftlichen Veröffentlichungen aus den Jahren 2006 bis 2014. Die Veröffentlichungen umfassen die Anwendungsbereiche Mimik-, Action-Unit-, Valenz- und Geschlechterklassifikation und beschränken sich auf statische, merkmalsbasierte Verfahren. Eine tabellarische Auflistung der einzelnen Publikationen findet sich im Anhang B.3. Wie im Schaubild 4.2 ersichtlich wird, unterscheidet man bei den merkmalsbasierten Verfahren zwischen *distanzbasierten*, *patchbasierten* und *globalen* Ansätzen. Das Diagramm 4.3 lässt verschiedene Trends für die Verwendung dieser Ansätze in der Literatur erkennen. Während in der ersten Hälfte des betrachteten Zeitraums die Distanzmerkmale dominieren, nimmt ab dem Jahr 2010 die Anzahl der patchbasierten Verfahren zu.

Der BU-3DFE-Datensatz

BU-3DFE (Binghamton University 3D Facial Expression) beschreibt eine frei verfügbare Sammlung von Gesichtsmodellen zum Zweck der Emotionserkennung. Im Detail umfasst sie 3D-Tiefendaten und RGB-Bilder von insgesamt 100 Personen, welche die sechs verschiedenen Basisemotionen[2], sowie ein Neutralgesicht mimisch nachahmen [Yin et al., 2006]. Für jedes Gesicht sind zudem die Positionen von 83 manuell gelabelten Landmarken enthalten.

In insgesamt 17 der 22 ausgewerteten FER-Publikationen wird der BU-3DFE-Datensatz zur Evaluation verwendet. Nähere Details dazu finden sich in der Tabelle B.3 im Anhang. Aufgrund des Bezugs zur Emotionserkennung ist der Datensatz jedoch nicht für den Einsatz in dieser Arbeit geeignet.

Durch die erstmalige Nutzung des BU-3DFE-Datensatzes stellt die Veröffentlichung von Wang et al. [J. Wang et al., 2006] einen wichtigen Anfangspunkt in der Reihe der bis 2014 folgenden Publikationen dar. Das darin beschriebene Verfahren ist krümmungs- und patchbasiert. Zuerst wird jedem Vertex des dreidimensionalen Gesichtsmodells eine von zwölf möglichen kategorischen Krümmungen zugeordnet. Beispiele hierfür sind *ravine* (*dt.* Schlucht) oder *ridge* (*dt.* Bergkamm). Anschließend wird das Gesicht auf Basis von 64 der insgesamt 83 verfügbaren Landmarken in sieben Regionen aufgeteilt und für jede Region ein Histogramm über die Krümmungskategorien gebildet.

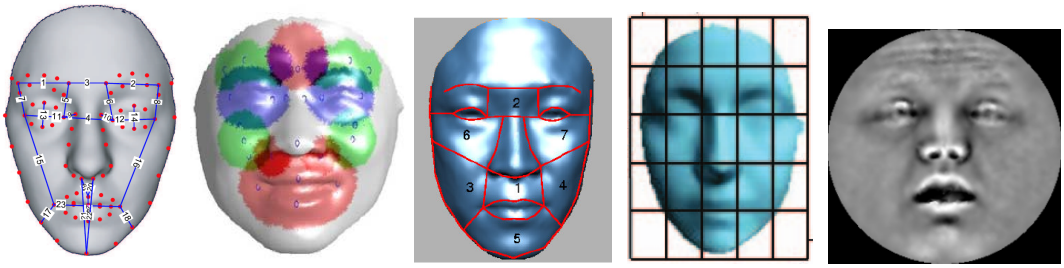
Distanzbasierte Ansätze

Obwohl der Ansatz von [J. Wang et al., 2006] patchbasiert ist, dominiert in den Folgejahren bis 2009 die distanzbasierte Vorgehensweise. Insgesamt acht der 28 ausgewerteten Publikationen sind dieser Kategorie zuzuordnen[3]. Sie stützen sich auf die Annahme, dass Mimikbewegungen Auswirkungen auf die euklidischen Abstände der Landmarken im Gesicht haben. Ein Lächeln beispielsweise bewirkt eine Dehnung der Lippen und somit eine Vergrößerung des Abstandes zwischen den Mundwinkeln.

Allerdings sind die absoluten Distanzen für den Einsatz als Merkmalsdeskriptoren ungeeignet, da sie von Individuum zu Individuum variieren und nicht altersunabhängig sind. Aus diesem Grund ist eine Normalisierung der Absolutwerte sinnvoll, beispielsweise anhand der Division durch ein invariantes Referenzmaß des Gesichts. Mögliche Referenzmaße sind die Gesichtsbreite oder der Abstand zwischen den Au-

[2]Die Basisemotionen sind Wut, Ekel, Angst, Freude, Traurigkeit und Überraschung.

[3]Distanzbasierte Publikationen: [Soyel und Demirel, 2007], [H. Tang und Thomas S. Huang, 2008], [Soyel und Demirel, 2008], [H. Tang und Thomas S Huang, 2008], [Srivastava und Roy, 2009], [X. Li et al., 2010], [C. Li und Soares, 2011], [Rabiu et al., 2012].



(a) (b) (c) (d) (e)

Abbildung 4.4.: **(a)** Der Verlauf der 24 in [H. Tang und Thomas S. Huang, 2008] verwendeten Distanzen. **(b)** Die in [Lemaire et al., 2011] zur Merkmalsextraktion ausgewählten radialen Patches. In ihrem Zentrum befinden sich die Landmarken. **(c)** Beispiel von [J. Wang et al., 2006] für die Unterteilung des Gesichts in Patches durch Verbinden der Landmarken mit geodätischen Linien. **(d)** Beispielabbildung aus [Bayramoglu et al., 2013] für die Unterteilung des Gesichts in Patches anhand eines regelmäßigen, landmarkenunabhängigen Gitters. **(e)** Globaler Ansatz von [Zeng et al., 2013]. Das Gesicht wird zur Merkmalsextraktion nicht in Regionen unterteilt. *(Die Abbildungen sind den jeweils angeführten Literaturquellen entnommen.)*

geninnenwinkeln ([Soyel und Demirel, 2007], [Soyel und Demirel, 2008], [X. Li et al., 2010], [C. Li und Soares, 2011], [Rabiu et al., 2012]). Eine weitere Variante stellt die Division durch die im Neutralgesicht gemessenen Abständen dar. Diese können korrespondierende, also durch dieselben Landmarken definierte, Distanzen sein oder eine Gruppen von festgelegten Normalisierungsmaßen (vgl. [P. Wang et al., 2007],[H. Tang und Thomas S. Huang, 2008], [H. Tang und Thomas S Huang, 2008]). In [Srivastava und Roy, 2009] werden Magnitude und Richtung des Versatzes zum Neutralgesicht bestimmt. Allerdings liegen nicht bei jedem Anwendungsszenario Aufnahmen der Neutralgesichter vor. So kann das Ausgangsgesicht bei Patienten mit Mimikdysfunktionen deutlich von einem Neutralgesicht im klassischen Sinne abweichen.

Ein weiteres Problem der Distanzmerkmale ist, dass sie die Analyse eines Gesichts auf wenige Einzelpunkte reduzieren. Dies ist insbesondere in Hinblick auf eine automatisierte Lokalisierung problematisch, da Ungenauigkeiten in der Landmarkenpositionierung einen starken Einfluss auf die zwischen den Landmarken gemessenen Distanzen haben können. Die Abbildung 4.4a zeigt beispielhaft den Verlauf der in [H. Tang und Thomas S. Huang, 2008] extrahierten Distanzen.

Die Anzahl der erforderlichen Landmarken variiert je nach Verfahren und liegt zwischen fünf und 83, wobei neun der zehn distanzbasierten Veröffentlichungen manu-

ell annotierte Landmarken einsetzen[4]. Nur der Ansatz von [P. Wang et al., 2007] stützt sich auf automatisiert lokalisierte Landmarken. Diese werden über ein *Active-Appearance-Model* (AAM) ermittelt [Cootes et al., 2001], welches zuvor iterativ an das zu analysierende Gesicht angepasst wurde. Während die nachfolgende Merkmalsextraktion sowohl auf 2D- als auch 3D-Daten basiert, erfolgt die AAM-Anpassung ausschließlich auf dem 2D-Intensitätsbild. Bei der Interpretation der Ergebnisse ist jedoch zu beachten, dass die Anzahl der zu untersuchenden Basisemotionen von sieben auf fünf reduziert wurde. Die Klassen *Neutral* und *Überraschung* fehlen. An dieser Stelle unterscheidet sich die Publikation von [P. Wang et al., 2007] von den anderen ausgewerteten Publikationen.

In einigen Publikationen werden die Distanzen, welche im Prinzip den Magnituden der euklidischen Landmarkenabstände entsprechen, um die Extraktion weiterer Maße ergänzt. Zu diesen zählen beispielsweise die Richtungen der Distanzvektoren [H. Tang und Thomas S Huang, 2008], die Winkel zwischen zwei Distanzvektoren ([C. Li und Soares, 2011],[Rabiu et al., 2012]), sowie Kombinationen beider Varianten [X. Li et al., 2010].

Patchbasierte Ansätze

Insgesamt sechzehn der 28 ausgewerteten Publikationen verwenden eine patchbasierte Vorgehensweise. In zwei Fällen werden zusätzlich Distanzmerkmale extrahiert, weshalb diese den gemischten Verfahren zuzuordnen sind ([P. Wang et al., 2007], [Bayramoglu et al., 2013]).

Im Unterschied zu den distanzbasierten Ansätzen, deren Merkmalsextraktion sich auf die Topologie der Landmarken beschränkt, werden bei den patchbasierten Verfahren Merkmale aus flächigen Arealen des Gesichts, den sogenannten Patches, extrahiert. Die *Festlegung der Arealgrenzen* variiert und lässt sich im Prinzip in drei Kategorien unterteilen (siehe Tab. 4.1). Bei der ersten Methode, welche in Abbildung 4.4b zu sehen ist, wird die Nachbarschaft einer Landmarke als Patch definiert. Die Begrenzung des Areals erfolgt zum Beispiel durch einen radialen Bereich um die Landmarke als Zentrum des Patches ([Maalej et al., 2010], [Maalej et al., 2011], [Lemaire et al., 2011], [Berretti et al., 2011]). Bei der zweiten Methode werden die Patch-Grenzen durch das Verbinden der Landmarken entlang der Oberfläche festgelegt, wie Abbildung 4.4c zeigt ([J. Wang et al., 2006], [P. Wang et al., 2007], [Y. Wang und Ma, 2012]). In insgesamt acht, und damit der Mehrheit der Publikationen, wird das Gesicht anhand eines regelmäßigen, $m \times n$ Felder umfassenden, Gitters unterteilt (siehe

[4]Die zehn Verfahren ergeben sich aus acht rein distanzbasierten Verfahren plus zwei gemischten Ansätzen.

Abb. 4.4d). Position und Größe des Gitters sind beispielsweise von einer Bounding-Box ableitbar, die über eine vorgeschaltete Gesichtsdetektion ermittelt werden kann. Eine zusätzliche Landmarkenlokalisierung ist somit nicht notwendig, was die Robustheit des Verfahrens erhöhen kann. Die Unabhängigkeit von den Landmarken führt jedoch auch dazu, dass die Zuordnung der einzelnen Felder zu bestimmten Gesichtsregionen nicht fix ist. Dies kann unter anderem dann von Nachteil sein, wenn sich durch bestimmte Bewegungen die Proportionen des Gesichts ändern, z.B. bei der Übung *AForm*.

Tabelle 4.1.: Einordnung der patchbasierten Verfahren nach der Art der Arealeingrenzung. Die grau unterlegten Publikationen sind gemischte Verfahren, d.h. sie verwenden zusätzlich Farbinformationen. Der patchbasierte Ansatz von [Xue et al., 2014] findet sich nicht in der Tabelle, da er eine Mischung aus allen drei Vorgehensweisen darstellt.

Regionenaufteilung	**Publikation**
Direkte LM-Nachbarschaft	[Maalej et al., 2010], [Maalej et al., 2011], [Lemaire et al., 2011], [Berretti et al., 2011]
LM-basierte Regionenaufteilung	[J. Wang et al., 2006], [P. Wang et al., 2007], [Y. Wang und Ma, 2012]
Regelmäßiges Gitter	[Sandbach et al., 2012a], [Sandbach et al., 2012b], [Broadbent et al., 2012], [Bayramoglu et al., 2013], [Lemaire et al., 2013], [Savran et al., 2013], [Huynh et al., 2013], [Y. Wang et al., 2013]

Nach dem die entsprechenden Regionen lokalisiert sind, beginnt die Extraktion der Merkmale. Die Extraktionsverfahren der ausgewerteten Publikationen sind relativ heterogen und lassen sich grob in vier Gruppen unterteilen. In Tabelle 4.2 findet sich eine Zuordnung der einzelnen Publikationen zu den verschiedenen *Extraktionsverfahren*. Die Mehrzahl der Verfahren nutzt die Krümmungsanalyse, um Informationen über den Verlauf der Gesichtsoberfläche zu gewinnen. Klassische, numerische Krümmungstypen sind unter anderem die mittlere Krümmung, die Gauß'sche Krümmung, der Shape-Index und die Curvedness. In [J. Wang et al., 2006] und [P. Wang et al., 2007] werden hingegen nicht-numerische Krümmungskategorien bestimmt. Die zweithäufigste Merkmalskategorie der patchbasierten Verfahren stellen Variationen der Local-Binary-Patterns (LBP) dar. Dieses Merkmalsextraktionsverfahren stammt ursprünglich aus der Texturanalyse von Grauwertdaten [Topi et al., 2000]. Bei der Extraktion von LBPs wird der Wert eines Pixels mit den Pixelwerten seiner Nachbarschaft verglichen und das extrahierte Muster als Binärzahl kodiert. Die dritthäufigste Kategorie der Merkmalsextraktion umfasst Ansätze, bei denen die Patches mit kor-

respondierenden Referenzpatches verglichen werden. Die Ähnlichkeit zwischen den Patches wird in Form eines Distanzmaßes beschrieben. Die verbleibenden drei Verfahren aus der Tabelle 4.2 lassen sich keiner der beschriebenen Kategorien zuordnen und sind daher unter *Sonstige* aufgeführt. In [Berretti et al., 2011] werden beispielsweise SIFT-Deskriptoren extrahiert.

Tabelle 4.2.: Vier verschiedene Merkmalskategorien, die auf der flächenbasierten Datenextraktion basieren. In der grau unterlegten Publikation wird eine Kombination aus Krümmungsanalyse und Local-Binary-Patterns eingesetzt.

Extraktionsverfahren	**Publikation**
Krümmungsanalyse	[J. Wang et al., 2006], [P. Wang et al., 2007], [Broadbent et al., 2012], [Lemaire et al., 2013], [Savran et al., 2013], [Y. Wang et al., 2013]
Local-Binary-Patterns	[Sandbach et al., 2012b], [Sandbach et al., 2012a], [Bayramoglu et al., 2013], [Huynh et al., 2013], [Y. Wang et al., 2013]
Patch-Abweichung	[Maalej et al., 2010], [Maalej et al., 2011], [Lemaire et al., 2011]
Sonstige	[Berretti et al., 2011], [Y. Wang und Ma, 2012], [Xue et al., 2014]

Gemischte Ansätze

Die gemischten Verfahren kombinieren die Extraktion von Distanz- und Patchmerkmalen ([P. Wang et al., 2007] , [Bayramoglu et al., 2013]). Die Besonderheit von [P. Wang et al., 2007] ist, dass die Distanzmerkmale nicht im dreidimensionalen Raum, sondern im zweidimensionalen Farbbild extrahiert werden.

Globale Ansätze

Die globalen Ansätze werden im Folgenden nicht weiter betrachtet, da sie Informationen aus dem gesamten Gesicht extrahieren und für eine gezielte Analyse einzelner Regionen nicht geeignet sind, wie Abbildung 4.4e zeigt ([Savran et al., 2010], [Yun und Guan, 2010], [Vretos et al., 2011], [Zeng et al., 2013]).

Weitere Ansätze

Bei der Recherche wurde nach bestem Wissen versucht, die tiefendatenbasierten Emotionserkennungsverfahren der Jahre 2000 bis 2014 systematisch abzudecken. Dennoch

stellen diese nur einen Ausschnitt der in Frage kommenden Verfahren dar, da eine umfassendere Übersicht und Analyse den Rahmen dieser Arbeit übersteigt.

Eine Sammlung von geeigneten Methoden zur punktwolkenbasierten Merkmalsextraktion findet sich in der *Point Cloud Library* (PCL) [Rusu und Cousins, 2011]. Die Anwendungsbereiche der von der PCL bereitgestellten Methoden beschränken sich nicht auf die Gesichtsanalyse, sondern können zum Beispiel auch auf die allgemeine Objekterkennung oder Erstellung von 3D-Modellen ausgeweitet werden. Einige der in dieser Arbeit eingesetzten Methoden, beispielsweise die Krümmungsanalyse, werden in ähnlicher Form von der PCL bereitgestellt. Die vorliegende Arbeit beschränkt sich jedoch auf eigene Implementierungen, um flexiblere Anpassungen der Methoden vornehmen zu können.

Die Literatur zur Personenerkennung wurde aus den in Abschnitt 4.1.1 beschriebenen Gründen weitestgehend ausgeklammert. Einige der Personenerkennungsverfahren lassen sich mittels entsprechender Änderungen dennoch an die Aufgabenstellung der Mimikerkennung anpassen. Das in [Chua et al., 2000] beschriebene Verfahren wurde für die Personenidentifizierung entwickelt, eignet sich jedoch allgemein zur Analyse von Oberflächenverläufen. Aus diesem Grund wird es in angepasster Form in die Evaluation einbezogen. Dazu wurde unter anderem die Merkmalsextraktion von der mimikunabhängigeren, starren Augenregion in den Mund- und Wangenbereich verlegt. Tiefergehende Details zur Implementierung finden sich im Unterkapitel 4.4.

4.1.3. Fazit und Wahl der Extraktionsverfahren

Im Rahmen der Literaturrecherche wurden systematisch 28 Veröffentlichungen mit Bezug zur tiefendatenbasierten Gesichtsanalyse ausgewertet. Eine detaillierte Auflistung der Veröffentlichungen findet sich im Anhang in der Tabelle B.3. Die beschriebenen Verfahren sind im Wesentlichen drei verschiedenen Extraktionskategorien zuzuordnen, welche als *distanzbasiert*, *patchbasiert* und *global* bezeichnet werden können. Die globale Merkmalsextraktion eignet sich nicht für das geplante Szenario, da eine Erweiterung zu einer regionenbezogenen Extraktion nicht möglich ist. Letztere ist jedoch für die Bestimmung von lokalem Feedback Voraussetzung, weshalb globale Verfahren nicht in die folgende Auswertung einbezogen werden.

Da insgesamt sechzehn der 28 ausgewerteten Veröffentlichungen auf einer patchbasierten Vorgehensweise gründen, werden zwei repräsentative Verfahren dieser Kategorie zur Evaluation ausgewählt. Diese umfassen die Krümmungsanalyse, welche mit sechs Publikationen die Mehrheit stellt, und die Histogramme orientierter Normalenvektoren (HON-Merkmale). Die distanzbasierte Merkmalsextraktion steht mit acht

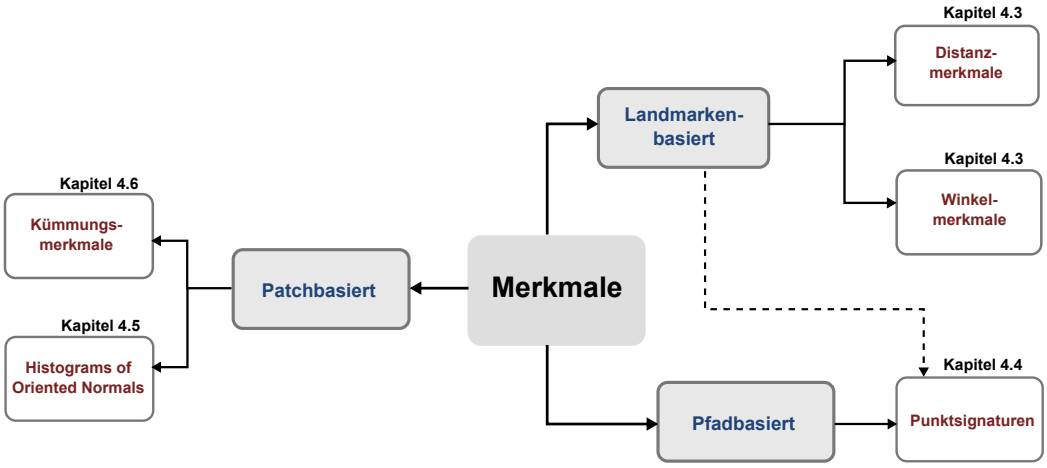


Abbildung 4.5.: Fünf Merkmalsextraktionsverfahren, die im Rahmen dieser Arbeit evaluiert werden (siehe Unterkapitel 4.3 bis 4.6).

Veröffentlichungen an zweiter Stelle, weshalb sie ebenfalls Teil der Evaluation ist. Als repräsentativer Vertreter wurde der Ansatz von [Rabiu et al., 2012] ausgewählt, da es sich bei diesem um den aktuellsten der acht ausgewerteten Ansätze handelt. Ergänzend zu den Distanzen werden die Winkel zwischen den Distanzvektoren extrahiert. In beiden Fällen ist eine vorgeschaltete Bestimmung von Landmarken notwendig, weshalb beide Vorgehensweisen als landmarkenbasierte Merkmalsextraktion bezeichnet werden. Ebenfalls in die Auswertung integriert werden die Punktsignaturen, welche ursprünglich für die Personenidentifizierung und Objekterkennung konzipiert wurden ([Chua und Jarvis, 1997], [Chua et al., 2000]). Sie erfordern lediglich die einer Zentrums- und einer Referenzlandmarke und werden aus Verläufen entlang der Gesichtsoberfläche extrahiert. Aus diesem Grund werden sie in dieser Arbeit nicht im klassischen Sinne den landmarkenorientierten Ansätzen zugeordnet. Alle gewählten Merkmalsextraktionsverfahren sind in der Abbildung 4.5 zusammengefasst, ergänzt um die ihnen zugeordneten Unterkapitel 4.3 bis 4.6.

Der Fokus dieser Arbeit liegt auf den Tiefenmerkmalen. Ergänzend wird im Unterkapitel 4.3 die landmarkenbasierte Merkmalsextraktion aus 2D-Bildern mit der Extraktion aus 3D-Punktwolken gegenübergestellt. Ein eingehenderer Vergleich von 3D- und 2D-Ansätzen übersteigt jedoch den Rahmen dieser Arbeit.

4.2. Allgemeines Testszenario

In diesem Abschnitt wird das allgemeine Testszenario der Merkmalsevaluationen anhand verschiedener Teilaspekte beschrieben. Merkmalsspezifische Anpassungen der

Vorgehensweise werden in den jeweiligen Merkmalskapiteln 4.3 bis 4.6 ausgeführt.

4.2.1. Datensatz und Vorverarbeitung

Der im Unterkapitel 2.3 beschriebene Datensatz von gesunden Personen bildet die Grundlage der folgenden Evaluationen. Er umfasst Aufnahmen von elf Personen während der Ausführung von insgesamt zwölf therapeutischen Fazialisübungen.

Das Ziel der Vorverarbeitung besteht in der einheitlichen räumlichen Registrierung aller 3D-Gesichtsmodelle. Dazu wird jede Punktwolke im ersten Schritt über eine Translation so verschoben, dass die Nasenspitze im Ursprung des Koordinatensystems liegt. Anschließend wird die Punktwolke an einer Referenzpunktwolke ausgerichtet. Die Ausrichtung erfolgt auf Basis des *Iterative-Closest-Point-Algorithmus*[5].

Falls für die Merkmalsextraktion eine Kenntnis der Landmarkenpositionen vonnöten ist, wird auf die manuell gelabelten Landmarken zurückgegriffen. Dadurch soll verhindert werden, dass sich Fehler und Ungenauigkeiten bei der automatisierten Lokalisierung auf die Ergebnisse der Merkmalsevaluation auswirken. Eine Auswertung der Merkmalsrobustheit gegenüber automatisiert geschätzten Landmarken findet sich im Abschnitt 5.5.1.

4.2.2. Klassifikationsbasiertes Evaluationsszenario

Im Rahmen der Evaluation wird ermittelt, inwieweit die extrahierten Merkmalsdeskriptoren übungsspezifische Informationen enthalten. Dazu werden die extrahierten Daten in Trainings- und Testdaten unterteilt und auf Basis der Trainingsdaten, sowie den ihnen zugeordneten Klassen, ein Klassifikationsmodell trainiert. Abschließend kann jedem Testbild eine, dem Modell nach wahrscheinlichste, Übungsklasse zugewiesen werden. Ein Vergleich von tatsächlicher und zugewiesener Klasse erlaubt eine Aussage über die Eignung der extrahierten Merkmalsdeskriptoren.

Um die Evaluationsbedingungen möglichst nah an die eines realen Übungsszenarios anzupassen, wird eine *personenbezogene Kreuzvalidierung* durchgeführt. Diese trägt der Tatsache Rechnung, dass während der Modellerstellung kein Wissen über den späteren Anwender der Übungsplattform vorliegt. Dementsprechend werden auch während der Evaluation sämtliche Bilder und Daten der Testperson aus dem Trainingsprozess ausgeschlossen.

[5]Die verwendete Funktion wurde dem MathWorks File Exchange entnommen, Autor: Jakob Wilm (http://www.mathworks.com/matlabcentral/fileexchange/27804-iterative-closest-point, letzter Zugriff: 18.09.2015).

Der Klassifikator selbst wird nicht variiert, da der Schwerpunkt des vorliegenden Kapitels auf den Extraktionsverfahren liegt. Bei allen fünf Merkmalstypen beschränkt sich die Evaluation auf den Einsatz von *Linear-Support-Vector-Machines* (Linear SVMs). Eine Auswertung mit verschiedenen Klassifikatoren findet sich im Unterkapitel 5.3. Die in den Experimenten eingesetzte SVM-Implementierung wird durch die *Library-for-Support-Vector-Machines* (LIBSVM) bereitgestellt [Chang und C.-J. Lin, 2011]. Diese bietet eine interne Anpassung für Mehrklassen-Probleme, da SVMs im Grunde binäre Klassifikatoren sind. Dabei werden $K(K-1)/2$ binäre Entscheider trainiert, wobei K die Anzahl der Klassen beschreibt. Da immer zwei Klassen einzeln gegeneinander getestet werden, bezeichnet man dies auch als one-versus-one Klassifikation. Jedes Ergebnis der binären Klassifikation stellt ein sogenanntes Voting (*dt.* Stimme) für eine bestimmte Klasse dar. Die Klasse mit den meisten Stimmen wird als finales Ergebnis ausgewählt.

Mit Hilfe des Strafparameters C (*engl.* penalty parameter) lässt sich festlegen, wie streng sich die Trenngrenzen der SVM an die Trainingsdaten anpassen sollen, d.h. in welchem Maße Falschklassifikationen beim Training des Klassifikationsmodells erlaubt sind. Dies kann unter anderem einen Einfluss auf die Generalisierfähigkeit des Klassifikatormodells haben. Der Parameter C wurde, analog zu der in [Hsu et al., 2003] beschriebenen Vorgehensweise, über eine Gridsuche bestimmt.

Zur Analyse der Relevanz einzelner Merkmalsvariablen wird die Mutual-Information (MI) ermittelt. Die Schätzung der Mutual-Information erfolgte unter Zuhilfenahme einer, von Erik Schaffernicht am Fachgebiet für Neuroinformatik und Kognitive Robotik (TU Ilmenau) entwickelten, Matlab-Toolbox (nähere Details zur MI siehe [Schaffernicht, 2012]).

4.2.3. Evaluationsmaße

Die gewählten Evaluationsmaße orientieren sich an den Evaluationsmaßen der, im Unterkapitel 4.1 vorgestellten, FER-Veröffentlichungen (siehe z.B. [J. Wang et al., 2006]). Sie dienen dazu, den Klassifikationserfolg zu quantifizieren und setzen sich im Wesentlichen aus den *mittleren Erkennungsraten* (MER) und den *Konfusionsmatrizen* zusammen. Beide werden in den folgenden Absätzen in Kurzform erläutert.

Ausgangspunkt sind ein balancierter Datensatz, sowie Testdaten einer p-ten Testperson, denen von einem gewählten Multiklassen-Klassifikator jeweils eine von K möglichen Übungsklassen zugeordnet wurden. Auf Basis dieser zugeordneten Klassen lassen sich die übungsspezifischen Recall-Raten $R_{p,k}$ schätzen. Die Recall-Rate

der Klasse $k = 1, ..., K$ ergibt sich aus:

$$R_{p,k} = \frac{TP_{p,k}}{TP_{p,k} + FN_{p,k}}. \tag{4.1}$$

Dabei beschreibt $TP_{p,k}$ die Anzahl der Richtig-Positiven-Entscheidungen und $FN_{p,k}$ die Summe aller Falsch-Negativen-Zuordnungen. Letztere fließen unabhängig von der zugeordneten Klasse k_{FNE} ein, wobei gilt $(k_{FNE} \in \{1, ..., K\}) \wedge (k_{FNE} \neq k)$.

Entsprechend der im vorhergehenden Abschnitt beschriebenen personenbezogenen Kreuzvalidierung fungiert jede der $P = 11$ Personen einmal als Testperson. Aus diesem Grund werden alle Ergebnisse P-mal wiederholt und über alle Personen gemittelt:

$$R_k = \frac{1}{P} \sum_{p=1}^{P} R_{p,k}. \tag{4.2}$$

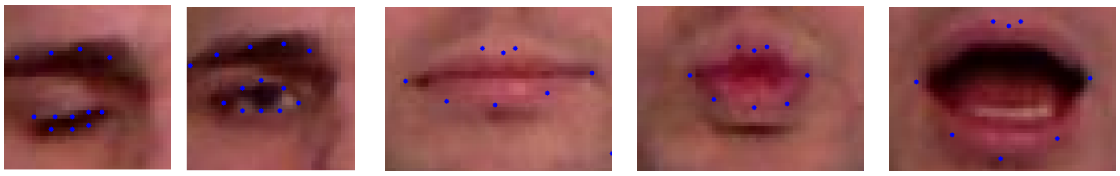
Aus dieser, über alle Testpersonen gemittelten, übungsspezifischen Recall-Rate R_k lässt sich als Endergebnis die mittlere Erkennungsrate $Recall_M$ (MER) berechnen [Sokolova und Lapalme, 2009]:

$$Recall_M = \frac{1}{K} \sum_{k=1}^{K} R_k. \tag{4.3}$$

Weiterführende Informationen zu Multiklassen-Evaluationsmaßen finden sich unter anderem in [Sokolova und Lapalme, 2009]. Die K übungsspezifischen Erkennungsraten R_k bilden die Diagonaleinträge der sogenannten *Konfusionsmatrix* (*engl.* confusion matrix). Die übrigen Matrixeinträge beinhalten die Vertauschungsraten zwischen den Observationen einer Referenzklasse k und einer Falsch-Positiven-Klasse k', mit $k \neq k'$. Beispielhafte Konfusionsmatrizen finden sich unter anderem in der Abbildung 4.10. Bei einer erfolgreichen Multiklassen-Klassifikation konzentrieren sich die Einträge mit den höchsten Beträgen auf der Hauptdiagonale der Konfusionsmatrix.

4.3. Distanz- und Winkelmerkmale

Den landmarkenbasierten Merkmalsextraktionsverfahren liegt die Annahme zu Grunde, dass sich Mimikbewegungen auf die räumlichen Relationen der Landmarken auswirken. Die Abbildungen 4.6a bis 4.6e verdeutlichen dies. In 4.1.2 wurden verschiedene landmarkenbasierte Verfahren vorgestellt. Zur Nachimplementierung wurde der aktuellste Ansatz [Rabiu et al., 2012] der ausgewerteten distanzmerkmalsbezogenen Publikationen ausgewählt. In Übereinstimmung zur beschriebenen Vorgehensweise werden neben den Distanzen auch die Winkel zwischen den landmarkenverbindenden



(a) (b) (c) (d) (e)

Abbildung 4.6.: Die Abbildungen (a) bis (e) verdeutlichen die durch die Mimikbewegungen hervorgerufenen Veränderungen der Landmarkenrelationen.

Distanzvektoren extrahiert und in die Analyse einbezogen. Das Unterkapitel gliedert sich in drei Teile, bestehend aus einer Vorstellung der Vorgehensweise und Ergebnisse von [Rabiu et al., 2012] (Abschn. 4.3.1), einer experimentellen Auswertung der eigenen Nachimplementierung (Abschn. 4.3.2), sowie einem abschließenden Fazit.

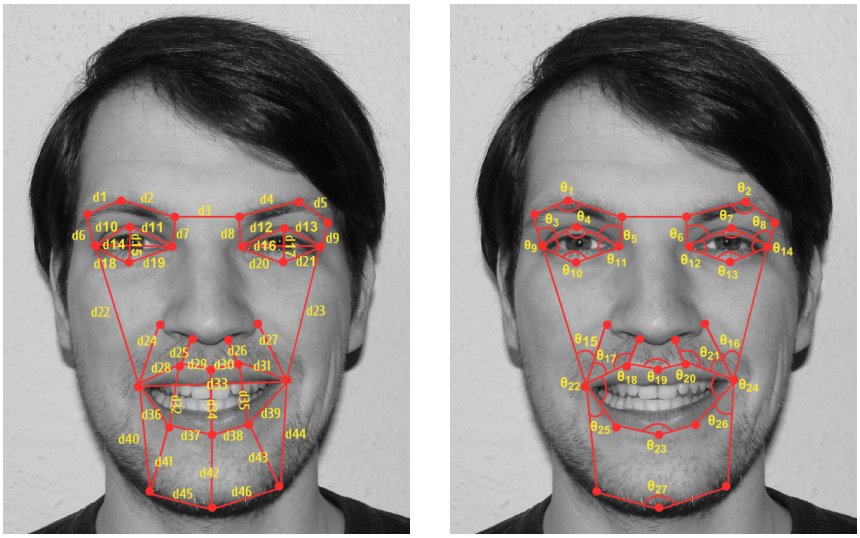
4.3.1. Distanzbasierte Merkmalsextraktion nach [Rabiu et al., 2012]

Das Ziel in [Rabiu et al., 2012] besteht in der automatisierten Erkennung des Neutralgesichts, sowie der sechs Basisemotionen Wut, Ekel, Angst, Freude, Traurigkeit und Überraschung. Die Grundlage der Experimente sind der eigens erstellte Bilddatensatz UPM-3DFE und der Referenzdatensatz BU-3DFE [Yin et al., 2006]. Beide enthalten 3D-Gesichtsmodelle mit manuell annotierten Landmarken im Bereich von Augen, Nase und Mund. Zur Extraktion der in [Rabiu et al., 2012] festgelegten Distanzen und Winkel werden pro Gesichtsmodell 29 Landmarken benötigt, im Folgenden gekennzeichnet durch $\beta_i = (x_i, y_i, z_i)$ mit $i \in \{1, ..., 29\}$. Um die Gesichtsmodelle einheitlich zu skalieren und beispielsweise unterschiedliche Kopfgrößen auszugleichen, werden alle Landmarkenvektoren durch eine Division mit der euklidischen Augeninnenwinkeldistanz normalisiert. Die erste Vorstufe der Distanz- und Winkelmerkmalsextraktion beinhaltet die Bestimmung von 46 euklidischen Basisdistanzen $d_{Basis} \in \{d1, ..., d46\}$, jeweils zwischen zwei Landmarken k und j:

$$||d|| = \sqrt{(x_k - x_j)^2 + (y_k - y_j)^2 + (z_k - z_j)^2} \tag{4.4}$$

Die Lage der Basisdistanzen innerhalb des Gesichts ist in Abbildung 4.7a zu sehen. Die Basisdistanzen werden im nächsten Schritt zu semantisch aussagekräftigeren Distanzmerkmalen δ_h, mit $h \in \{1, ...,16\}$, kombiniert. Diese neuen Distanzen und die von ihnen repräsentierte Gesichtsbewegung sind in Tabelle 4.3 aufgelistet.

Der zweite Teil der zu extrahierenden Merkmale umfasst die Winkel θ_l, mit $l \in$



(a) (b)

Abbildung 4.7.: Merkmalsextraktion nach [Rabiu et al., 2012]. **(a)** 46 Basisdistanzen der 16 Distanzmerkmale **(b)** 27 Winkelmerkmale

Gesichtsbewegung	**Berechnung**
Dehnung der Augenbrauen	$\delta_1 = d1 + d2, \quad \delta_2 = d4 + d5$
Runzeln der Augenbrauen	$\delta_3 = d3$
Vertikale Verschiebung der Innen- bzw. Außenseite der Augenbrauen	$\delta_4 = \frac{1}{2}(d7 + d8), \quad \delta_5 = \frac{1}{2}(d6 + d9)$
Öffnung der Augen	$\delta_6 = d15, \quad \delta_7 = d17$
Durchschnittliche Augenbreite	$\delta_8 = \frac{1}{2}(d14 + d16)$
Vertikale Öffnung des Mundes	$\delta_9 = \frac{1}{3}(d32 + d34 + d35)$
Positionverschiebungen der Mundwinkel	$\delta_{10} = \frac{1}{2}(d22 + d23),$ $\delta_{11} = \frac{1}{2}(d24 + d27), \quad \delta_{14} = d33$
Verschiebung und Dehnung der Oberlippe	$\delta_{12} = \frac{1}{2}(d25 + d26),$ $\delta_{13} = \frac{1}{4}(d28 + d29 + d30 + d31)$
Dehnung und Verschiebung der Unterlippe	$\delta_{15} = \frac{1}{4}(d36 + d37 + d38 + d39),$ $\delta_{16} = \frac{1}{5}(d40 + d41 + d42 + d43 + d44)$

Tabelle 4.3.: Berechnung der Distanzmerkmale nach [Rabiu et al., 2012].

$\{1, ...,27\}$, zwischen den Basisdistanzvektoren. Sie sind in Abbildung 4.7b visualisiert und werden definiert durch:

$$\theta_l = arccos\left(\frac{s_m \cdot s_n}{\|s_m\| \cdot \|s_n\|}\right), \tag{4.5}$$

wobei s_m und s_n die winkeleingrenzenden dreidimensionalen Vektoren darstellen.

Die berechneten Distanzen δ_h und Winkel θ_l werden zum Schluss zu einem 43 Dimensionen umfassenden Merkmalsvektor $\mathbf{M}$ zusammengefasst:

$$\mathbf{M} = [\delta_1, ..., \delta_{16}, \theta_1, ..., \theta_{27}]\,. \tag{4.6}$$

Um für das Training des Klassifikators ausschließlich relevante Merkmale beizubehalten und die redundanten zu entfernen, wird in [Rabiu et al., 2012] der mRMR-Algorithmus (Maximum Relevance Minimum Redundancy) auf die extrahierten Daten angewendet. Die Bewertung der einzelnen Dimensionen hinsichtlich Redundanz und Relevanz basiert dabei auf der Mutual-Information. Mit dem verbliebenen Merkmalssatz wird eine Mehrklassen-SVM trainiert. Diese erzielt auf dem BU-3DFE Datensatz eine mittlere Erkennungsrate von 92,2 %, für die Kombination von BU-3DFE und UPM-3DFE eine Rate von 88,9 %.

4.3.2. Experimentelle Untersuchung

Im Rahmen dieser Arbeit wurde der Ansatz von [Rabiu et al., 2012] nachimplementiert und unter Einsatz des therapeutischen Übungsdatensatzes evaluiert. Die folgende experimentelle Auswertung gliedert sich in drei Abschnitte. Diese umfassen

- eine allgemeine Ergebnisübersicht,
- eine detaillierte merkmals- und übungsspezifische Analyse der Ergebnisse, sowie
- einen Vergleich mit den korrespondierenden zweidimensionalen Distanz- und Winkelmerkmalen.

Die experimentelle Vorgehensweise und Bestimmung der Bewertungsmaße entspricht dabei dem im Unterkapitel 4.2 beschriebenen, allgemeinen Testszenario.

Allgemeine Ergebnisübersicht

Bei der Klassifikation in zwölf Übungsklassen wurde, unter Verwendung aller 43 Merkmalsdimensionen, eine mittlere Erkennungsrate (MER) von $Recall_M = 60{,}62\,\%$ er-

reicht. Die elf personenspezifischen, über alle Übungsklassen gemittelten, Erkennungsraten liegen zwischen 44,05 % und 77,18 %. Im Gegensatz dazu sind die zwölf übungsspezifischen Erkennungsraten R_k deutlich stärker gestreut und bewegen sich zwischen 14,29 % und 100 %[6].

Durch die Beschränkung auf eine ausgewählte Untermenge der 43 Merkmalsdimensionen lässt sich die MER von 60,62 % auf 64,82 % verbessern. Details zur Merkmalsselektion finden sich im folgenden Abschnitt. Die in dieser Arbeit erzielten Erkennungsraten liegen dennoch unter den in [Rabiu et al., 2012] geschilderten Referenzergebnissen, welche, in Abhängigkeit vom getesteten Datensatz, 88,9 % beziehungsweise 92,2 % betragen. Diese Unterschiede haben mehrere Gründe. Zum einen basieren die Datensammlungen auf unterschiedlichen Aufnahmesystemen und haben somit auch abweichende Tiefenauflösungen. Zum anderen unterscheiden sich sowohl Anwendungsszenario als auch Klassenanzahl. Die Erkennung von sieben Emotionsklassen im Referenzverfahren steht der Erkennung von zwölf Übungsklassen in dieser Arbeit gegenüber.

Übungs- und merkmalsbezogene Auswertung

Zur ausführlicheren Auswertung der erzielten Ergebnisse werden zwei verschiedene Werkzeuge der Datenanalyse eingesetzt. Die Konfusionsmatrix, welche die Vertauschungen zwischen den Klassen visualisiert, und die Mutual-Information (MI), welche die Beziehung zwischen den Merkmalsvariablen (MV) und der Übungsklassenzielvariable verdeutlicht. Weiterführende Informationen zu beiden wurden bereits im Unterkapitel 4.2 gegeben.

Für alle Personen wurde die Mutual-Information zwischen den 43 Merkmalsvariablen und der Übungsklassenzielvariable geschätzt. Die über alle Personen gemittelten Schätzwerte sind in der Abbildung 4.8 in einem Säulendiagramm visualisiert. Das Diagramm verdeutlicht, dass die der Mundregion zugeordneten Merkmalsvariablen einen größeren Zusammenhang zur ausgeführten Übung aufweisen, als die MV der Augenbrauen- und Augenregion. Dieses Ergebnis deckt sich mit der Tatsache, dass die Übungen vorwiegend der Mobilisierung von Mund- und Wangenmuskulatur dienen und in diesen Regionen entsprechend die größten Mimikbewegungen stattfinden. In den Abbildungen 4.9a und 4.9b sind ausgewählte Distanzen und Winkel eingezeichnet. Insgesamt findet sich bei den Merkmalen, welche die horizontale und vertikale Öffnung des Mundes beschreiben, der größte Zusammenhang zur ausgeführten Übung. Inner-

[6] Übungsspezifische Erkennungsraten: *Augen* 100 %, *Kuss* 64,7 %, *Breit* 83,12 %, *AForm* 98,7 %, *OForm* 39,67 %, *IForm* 86,15 %, *UForm* 35,64 %, *BoxenLi* 53,17 %, *BoxenRe* 14,29 %, *Wangen* 62,19 %, *WangeLi* 32,95 %, *WangeRe* 56,82 %.

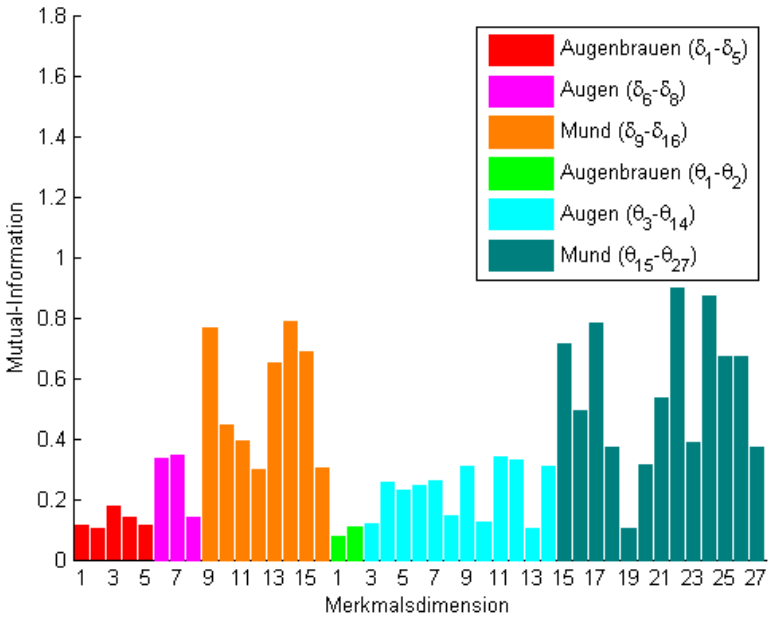
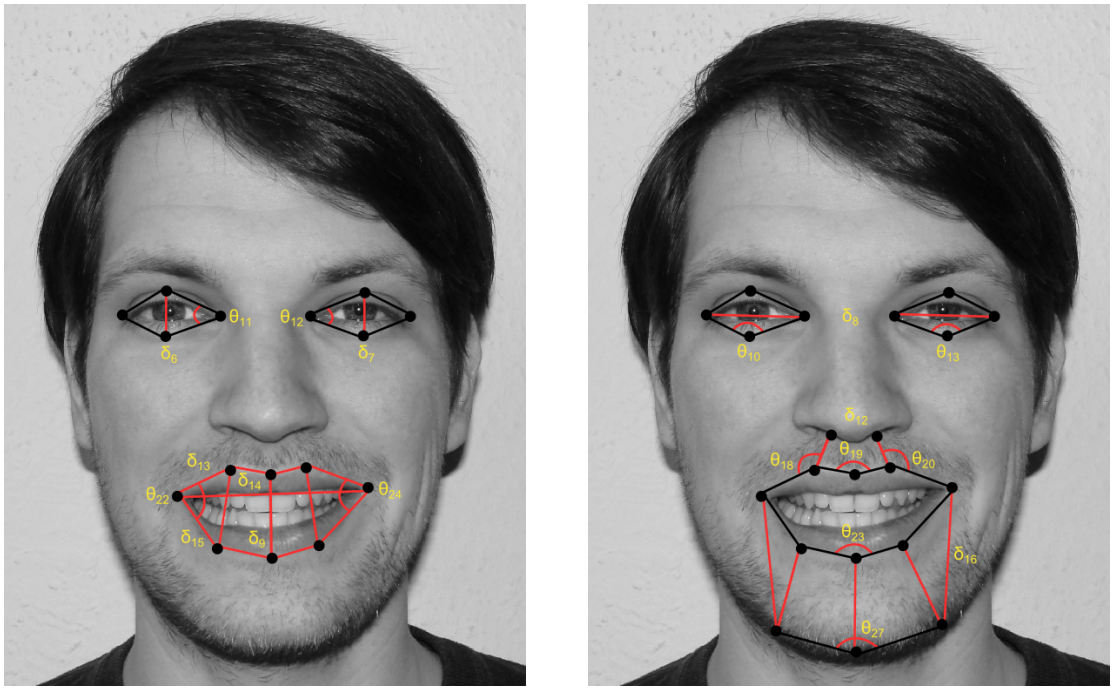


Abbildung 4.8.: Die geschätzte mittlere Mutual-Information zwischen den 43 Merkmalsvariablen und der Übungsklassenzielvariable. Die ersten 16 Säulen beziehen sich auf die Distanzen δ_h, die 27 folgenden auf die Winkel θ_l. Die Farbe ordnet die Merkmale den einzelnen Regionen des Gesichtes zu.

halb der Augenregion stellt die Öffnung des oberen Augenlids das aussagekräftigste Charakteristikum dar. Im Vergleich zu den Merkmalen der Mundregion ist es jedoch nur von mittlerer Relevanz. Merkmalsvariablen, deren Werte nur geringe Zusammenhänge zu den zugeordneten Übungen aufweisen sind beispielsweise solche, welche die Distanz und Lage der Mundlandmarken zu Nase oder Kinn beschreiben. Die Extraktion der relativ konstanten, mittleren Augenbreite δ_8 trägt erwartungsgemäß kaum zu einer korrekten Klassifikation bei.

Betrachtet man die Distanz- und Winkelmerkmale als zwei getrennte Gruppen und berechnet die mittlere MI über alle 16 bzw. 27 Dimensionen, dann ergeben sich MI-Werte von 0,362 und 0,375. Separate Klassifikationen bestätigen den Relevanzvorsprung der Winkelmerkmale. Deren Einsatz erzielte mit 64,16 % eine deutlich höhere MER als die distanzmerkmalsbasierte Klassifikation, welche eine Rate von 51,08 % ergab. Bei der Interpretation der Ergebnisse ist jedoch auch zu beachten, dass die Winkelmerkmale eine höhere Dimensionsanzahl aufweisen.

Vergleicht man die gruppenbezogenen Ergebnisse mit der MER des kombinierten Merkmalssatzes (60,62 %), dann zeigt sich einerseits eine deutliche Verschlechterung durch das Ausschließen der Winkelmerkmalsdeskriptoren (51,08 %) und andererseits eine leichte Verbesserung durch das Ausschließen der Distanzmerkmalsdeskriptoren



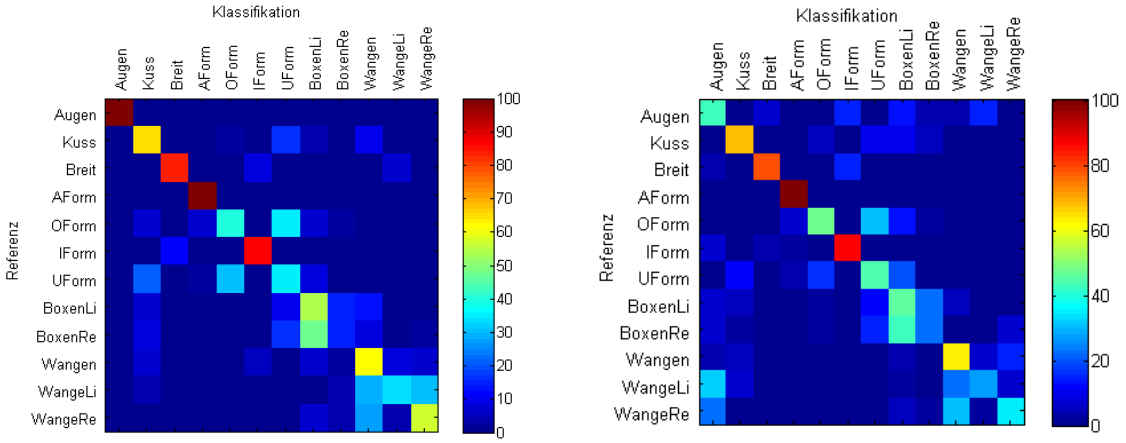
(a) (b)

Abbildung 4.9.: Eine Auswahl der Distanz- und Winkelmerkmale, die eine vergleichsweise **(a)** hohe bzw. **(b)** geringe MI mit der Zielvariable aufweisen. Zu beachten ist, dass die MI-Werte der besten Augenmerkmale – δ_6, δ_7, θ_{11}, θ_{12} – nur vergleichbar mit den schwächeren Mundmerkmalen, z.B. δ_{12}, sind. Die Augenbrauenmerkmale weisen alle eine sehr geringe MI auf und wurden daher nicht eingezeichnet. Weiterführende Erläuterungen zu den einzelnen Distanzen und Winkel finden sich in Tabelle 4.3 und Abbildung 4.7b.

(64,16 %). Dies deutet auf, zumindest in Teilen, redundante oder wenig relevante Distanzmerkmalsdimensionen hin. Ergänzt man die Winkelmerkmale nur um die vier Distanzmerkmale δ_9, δ_{13}, δ_{14} und δ_{15} mit der höchsten MI, lässt sich eine leichte Verbesserung um 0,58 Prozentpunkte auf 64,74 % erreichen.

Obwohl die Mundmerkmalsvariablen $(\delta_9 - \delta_{16}, \theta_{15} - \theta_{27})$ insgesamt die höchsten MI-Werte aufweisen, führt eine Beschränkung auf diese allein zu einer vergleichsweise geringen mittleren Erkennungsrate von 55,24 %. Dies lässt sich durch einen Blick auf die Konfusionsmatrizen in den Abbildungen 4.10a und 4.10b erklären. Da keine Informationen aus der Augenregion in den Klassifikationsprozess einfließen, verschlechtert sich die übungsspezifische Erkennungsrate der Klasse *Augen* von 100 % deutlich auf 43,34 %. Die Abbildungen 4.6a und 4.6b verdeutlichen die Unterschiede der Landmarkenanordnung zwischen der *Augen*-Übung und den anderen Übungen.

Eine Ergänzung der Mundmerkmale um die sechs aussagekräftigsten Deskriptoren



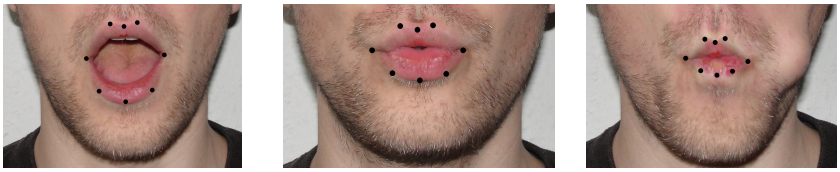
(a) (b)

Abbildung 4.10.: Konfusionsmatrizen für die Klassifikation der zwölf therapeutischen Übungen mit verschiedenen Konfigurationen. Die MER ergibt sich als Mittelwert der Diagonaleinträge. **(a)** Verwendung aller Merkmalsdimensionen δ_h und θ_l (MER 60,62 %). **(b)** Beschränkung auf Mundmerkmale (MER 55,24 %).

der Augenregion $(\delta_6, \delta_7, \theta_9, \theta_{11}, \theta_{12}, \theta_{14})$ verbessert die MER von 55,24 % auf 64,82 % und erzielt das bisher beste Ergebnis. Der Unterschied zur Variante mit der zweitbesten MER (64,74 %) ist sehr gering, es werden jedoch vier Merkmalsdimensionen weniger benötigt.

Anhand der Konfusionsmatrizen wird zudem deutlich, dass die Vertauschungen zwischen den Klassen für die Übungen *BoxenLi*, *BoxenRe*, *Wangen*, *WangeLi* und *WangeRe* besonders hoch ausgeprägt sind. Hier zeigt sich ein großer Nachteil des Einsatzes von Distanz- und Winkelmerkmalen. Aufgrund der landmarkenbasierten Merkmalsextraktion, werden Informationen über die Gesichtsoberfläche in homogenen, landmarkenarmen Regionen nicht in den Trainings- und Klassifikationsprozess einbezogen. Dadurch werden dem Klassifikator, möglicherweise relevante, Informationen vorenthalten, insbesondere bei Übungen, die sich weniger durch eine Verschiebung der Landmarken als durch eine Änderung der Gesichtsoberfläche, zum Beispiel im Wangenbereich, auszeichnen (siehe Abbildung 4.11).

Die bisherigen Ergebnisse basieren auf einer linearen Multiklassen-SVM mit einem Strafparameter $C = 1$, dem Default-Wert der LIBSVM. Im nächsten Schritt wird über eine Grid-Suche und eine 5-fach-Kreuzvalidierung auf den Trainingsdaten ein Wert für C bestimmt und die Unbalanciertheit der Daten ausgeglichen (siehe Unterkapitel 4.2), was in einer MER von 63,64 % resultiert (Abb. 4.12). Diese Vorgehensweise wird bei den Evaluationen der anderen Merkmalstypen beibehalten.



(a) (b) (c)

Abbildung 4.11.: **(a)** Die Ausführung der Übung *AForm* führt zu einem charakteristischen Muster der Mundlandmarken. Dies resultiert in einer hohen übungsspezifischen Erkennungsrate von 98,86 %. **(b)**–**(c)** Die übungspezifischen Erkennungsraten der Klassen *UForm* und *BoxenLi* liegen bei 50,46 % bzw. 48,79 %. Obwohl sich beide Übungen in den Wangenregionen optisch deutlich unterscheiden, werden 12,31 % der *UForm*-Gesichter der Klasse *BoxenLi* zugeordnet und 13,64 % der *BoxenLi*-Gesichter der Klasse *UForm*. Dies gründet darauf, dass die Distanz- und Winkelmerkmale landmarkenbasiert sind und keine direkte Information aus der Wangenoberfläche extrahieren.

Intensitätsbildbasierte Merkmalsextraktion

Um den Beitrag der Tiefeninformation beurteilen zu können, wurde in dieser Arbeit zusätzlich ein Referenzverfahren implementiert, bei welchem die Distanzen und Winkel aus einem zweidimensionalen (Farb-)Bild extrahiert werden. Die Datenbasis besteht nicht mehr aus einer dreidimensionalen Punktwolke, sondern aus n Pixel-Landmarken $p_i = (x, y)$, mit $i = (1, ..., n)$. Folglich sind die Distanzen nun in Pixeln und nicht in Metern angegeben.

Diese Loslösung von metrischen Einheiten hat jedoch Nachteile. So ist die Abbildungsgröße des Gesichts in einem Farbbild von der Kamera-Objekt-Distanz abhängig, wodurch die Abstände innerhalb des Gesichts bildübergreifend variieren können. Eine Normalisierung der Distanzen mit dem Augeninnenwinkel-Abstand gleicht dies aus und führt zu dimensionslosen Merkmalswerten, die mit denen der normalisierten 3D-Abstände vergleichbar sind.

Es wurde mit zwei verschiedenen Experimentkonstellationen getestet:

- Die Klassifikation mit allen 43 Distanz- und Winkelmerkmalsdeskriptoren, einem Default-Strafparameter $C = 1$ ohne Ausgleich eventueller Datenunbalanciertheit resultiert in einer mittleren Erkennungsrate von 53,24 %. Dies entspricht einer deutlichen Verschlechterung gegenüber der Klassifikation mit Tiefenmerkmalen (60,62 %).

- Die Klassifikation mit allen Mundmerkmalen und sechs Augenmerkmalen ($\delta_6, \delta_7, \theta_9, \theta_{11}, \theta_{12}, \theta_{14}$). Der Strafparameter C wird über eine 5-fache Kreuzva-

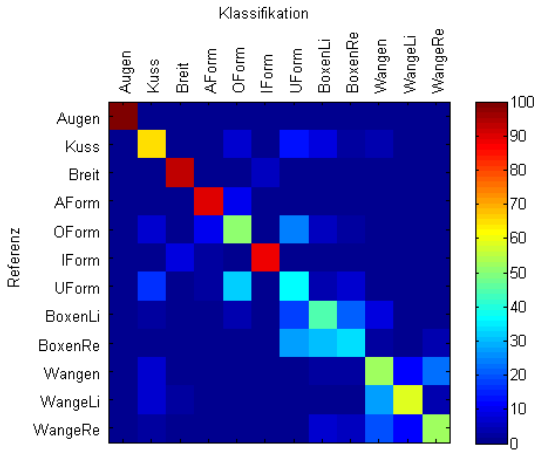


Abbildung 4.12.: Konfusionsmatrix der Klassifikation mit allen 27 Mundmerkmalsdimensionen und sechs Augenmerkmalsdimensionen mit der höchsten MI $(\delta_6, \delta_7, \theta_9, \theta_{11}, \theta_{12}, \theta_{14})$. Zusätzlich Ausgleich der Datenunbalanciertheit und experimentelle Bestimmung des Strafparameters C. Die mittlere Erkennungsrate beträgt MER 63,64 %.

lidierung aus den Trainingsdaten geschätzt und die Unbalanciertheit über klassenspezifische Wichtungsparameter ausgeglichen. Die erzielte MER von 54,95 % ist geringer als das Ergebnis bei Einsatz der Tiefenmerkmale (63,64 %).

Beide Resultate zeigen, dass die Integration von Tiefeninformationen bei der automatisierten Unterscheidung von therapeutischen Fazialisübungen förderlich ist.

4.3.3. Fazit

Wie in der Abbildung 4.3 ersichtlich wurde, dominierten zu Beginn des Untersuchungszeitraums distanzbasierte Ansätze die Gesichtsanalyse. Die entsprechende Merkmalsextraktion erfordert jedoch eine vorgeschaltete Lokalisierung der Landmarken, was im Hinblick auf ein automatisiertes Szenario nicht trivial ist und eine zusätzliche Fehlerquelle darstellen kann. Aus diesem Grund rückten in den Folgejahren patchbasierte und globale Verfahren in den Vordergrund.

Zur experimentellen Evaluation der Distanzmerkmalsdeskriptoren wurde das chronologisch aktuellste Verfahren des Auswertungszeitraums ausgewählt und nachimplementiert [Rabiu et al., 2012]. Das Verfahren kombiniert Distanz- und Winkelmerkmalsdeskriptoren, weshalb diese ebenfalls in die Analyse einbezogen wurden. Bei der Auswertung der mittleren Erkennungsraten zeigte sich, dass die winkelbasierte Klassifikation mit 64,16 % bessere Ergebnisse erzielt als die Klassifikation auf Basis der Distanzmerkmalsdeskriptoren (51,08 %) oder der Kombination beider Merkmalstypen

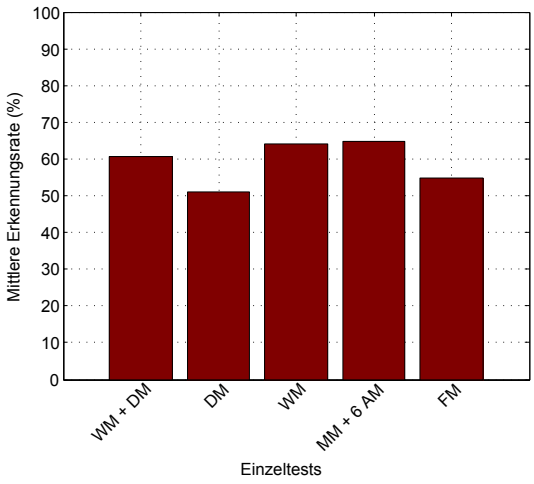


Abbildung 4.13.: Mittlere Erkennungsraten verschiedener Merkmalskonfigurationen. Abkürzungen: Winkelmerkmale (WM), Distanzmerkmale (DM), Mundmerkmale (MM), Augenmerkmale (AM), Farbmerkmale (FM)

(60,62 %). Das Balkendiagramm in der Abbildung 4.13 fasst diese und weitere Ergebnisse zusammen. Im Folgenden dient die MER von 63,64 % als Referenzwert, da sie, hinsichtlich der Parametersuche und der Datenanpassung, dem in Unterkapitel 4.2 beschriebenen Standard-Testszenario entspricht.

Bei der Analyse der Konfusionsmatrizen offenbart sich die größte Schwachstelle der landmarkenbasierten Vorgehensweise (siehe u.a. Abb. 4.12). Da keine Informationen aus landmarkenarmen Gesichtsregionen in den Trainings- und Klassifikationsprozess einfließen, weisen z.B. die Übungen *BoxenLi* und *UForm* hohe Vertauschungsraten auf, obwohl sie sich im Oberflächenverlauf der Wangenregionen deutlich unterscheiden.

4.4. Punktsignaturen

Die Punktsignaturen stellen in dieser Arbeit einen Sonderfall dar, da sie ursprünglich in ein Verfahren zur mimikunabhängigen Personenerkennung eingebettet waren [Chua et al., 2000]. Damit heben sie sich von den anderen Verfahren ab, die eine personenunabhängige Erkennung von Gesichtsausdrücken zum Ziel haben. Während sich die Punktsignaturextraktion in [Chua et al., 2000] auf starre Regionen des Gesichts, wie Nase, Augenhöhlen und Stirn konzentriert, wird die Merkmalsextraktion in dieser Arbeit auf die dynamische Mund- und Wangenregion verlagert.

Der erste Abschnitt 4.4.1 beschreibt diese adaptierte Variante der Punktsignaturextraktion. Die dieser Arbeit zugrunde liegende Implementierung des Verfahrens ist eine veränderte und erweiterte Version der von Birant Sibel Olgay im Rahmen ihrer

Masterarbeit umgesetzten Punktsignaturimplementierung [Olgay, 2012][7]. Im Abschnitt 4.4.2 folgt eine experimentelle Auswertung, basierend auf dem im Unterkapitel 4.2 vorgestellten Testszenario. Das abschließende Fazit fasst die wichtigsten Ergebnisse zusammen.

4.4.1. Theorie und Implementierung der Punktsignatur-Adaption

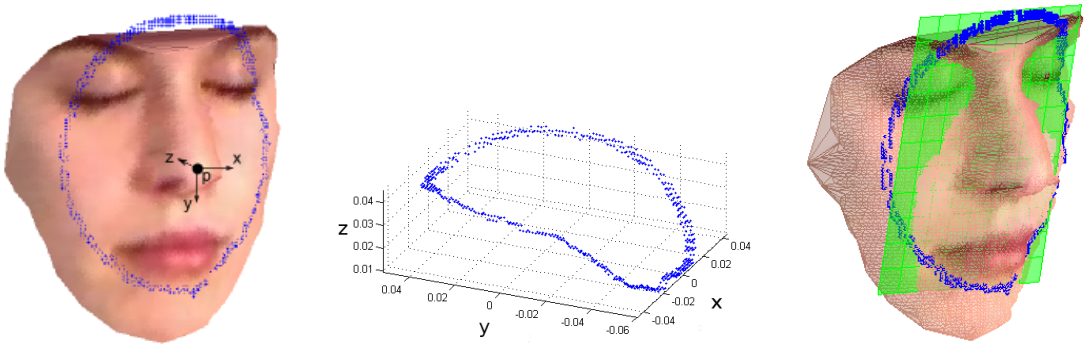
Eine Punktsignatur entspricht einer Raumkurve, die entlang einer Oberfläche verläuft. Verändert sich die Oberfläche, wirkt sich dies auch auf den Verlauf der Punktsignatur aus. In dieser Arbeit wird der Verlauf der Punktsignatur durch diskrete 3D-Punkte angenähert. Um den Punktsignaturverlauf in einen translations- und rotationsinvarianten Merkmalsdeskriptor überführen zu können, wird ein relatives Referenzsystem benötigt. Die direkte Extraktion der 3D-Weltkoordinaten (x, y, z) ist hingegen ungeeignet, da die Lage der Gesichtspunktwolke, und somit auch der Punktsignatur, im Koordinatensystem variabel ist. Die Extraktion von Punktsignaturen unter Verwendung einer relativen Referenzebene wird in den folgenden Absätzen detaillierter beschrieben.

Eine Punktsignatur zeichnet sich im Wesentlichen durch ein Punktsignaturzentrum p und einen Radius r aus. Die Nasenspitze dient in dieser Arbeit als Signaturzentrum p, da sie mittig im Gesicht liegt und aufgrund ihrer spezifischen Krümmung robust detektierbar ist. Dies ist insbesondere in Hinblick auf eine Automatisierung der Merkmalsextraktion von Vorteil. Die Punktsignatur verläuft, wie bereits erwähnt, entlang der Gesichtsoberfläche, wobei es sich bei letzterer genauer betrachtet um eine Ansammlung von diskreten Punkten handelt. Zu Beginn werden die Punkte der Oberfläche bestimmt, die einen euklidischen Abstand von $r \pm \epsilon$, zum Zentrum p haben. Die Ergänzung von r um einen Fehlerterm ϵ ist notwendig, da bei diskreten Daten das Auftreten exakter euklidischer Distanzen nicht gewährleistet ist. In dieser Arbeit wurde ϵ ein Wert von 2 mm zugewiesen[8]. Die gefundenen Punkte ergeben eine dreidimensionale Kurve C. Die Abbildungen 4.14a und 4.14b visualisieren ihren Verlauf entlang der Gesichtsoberfläche beziehungsweise freistehend im Koordinatensystem.

Um den Verlauf von C beschreiben zu können und ihn in einen translations- und rotationsinvarianten Merkmalsdeskriptor zu überführen, ist eine relative Referenzebene notwendig. Dazu wird mit Hilfe der Singulärwertzerlegung eine Ebene E in die diskre-

[7] Diese Masterarbeit wurde von der Autorin im Rahmen dieser Dissertation betreut.

[8] Der Fehlerterm ließe sich durch das Fitten einer Oberfläche mit anschließender Interpolation ersetzen. Dem stünde jedoch ein höherer Rechenaufwand gegenüber, da sich die aktuelle Implementierung die Vorteile der Vektorisierung zu Nutze macht. Details dazu finden sich in `http://de.mathworks.com/help/matlab/matlab_prog/vectorization.html`, letzter Zugriff: 24.09.2015.

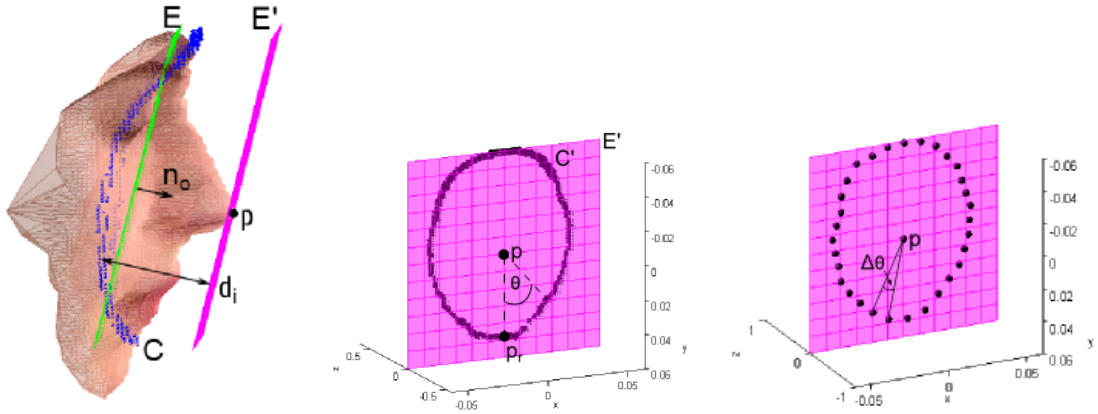


(a) (b) (c)

Abbildung 4.14.: **(a)** Die blau-markierten, diskreten Punkte weisen einen euklidischen Abstand von $r \pm \epsilon$ zum Referenzpunkt p auf. Sie bilden eine dreidimensionale Raumkurve C. **(b)** Lage der Raumkurve C im dreidimensionalen Koordinatensystem. **(c)** Gefittete Ebene E.

ten Punkte der Kurve C gefittet (siehe Abb. 4.14c). Details zu dieser Vorgehensweise finden sich im Anhang A.2.

Anschließend wird die Ebene in Richtung ihres Normalenvektors n_0 verschoben, bis sie den Zentrumspunkt p schneidet (siehe Abb. 4.15a). Für jeden der M diskreten Punkte von C wird der euklidische Abstand d_i, mit $i = 1, ..., M$, zu der aus der Verschiebung resultierenden Ebene E' bestimmt. Eine Verschiebung der Punkte um d_i in Richtung des Normalenvektors n_0 ergibt eine zweidimensionale Projektion C' auf die Ebene E'. Diese ist in Abbildung 4.15b zu sehen. Die Elemente von C' bestehen aus den 2D-Koordinaten (x, y), da $z = 0$ gilt. Jedem Punkt von C' ist somit ein euklidischer Abstand d_i zugeordnet. Diese Abstände bilden die Ausgangselemente der Punktsignaturmerkmalsdeskriptoren. Um sie in eine Vektorform zu überführen, werden die einzelnen Punkte von C' von kartesischen Koordinaten (x, y) in Polarkoordinaten (ρ, θ) umgewandelt. Im nächsten Schritt werden die Abstände d_i für $\theta = [-180°; +180°]$ ausgelesen (siehe Schaubild 4.16a). Damit eine Rotationsinvarianz des Deskriptors gewährleistet ist, muss $\theta = 0$ durch einen charakteristischen, wiedererkennbaren Referenzpunkt p_r definiert sein. Die Länge des vorliegenden Merkmalsvektors ist abhängig von der Anzahl der Punkte in C. Um den Deskriptor auf eine bild- und signaturübergreifende einheitliche Länge zu bringen, wird ein Abtastintervall $\Delta\theta$ festgelegt. Die Abbildung 4.15c zeigt die auf 32 Elemente reduzierte Kurve C' für ein Abtastintervall $\Delta\theta = 11{,}6°$. Das korrespondierende Schaubild ist in der Abbildung 4.16b zu sehen. Die Abstandswerte $d(\theta)$ der Abtastwinkel werden über eine lineare Interpolation aus den Abständen d_i ermittelt. Die von $\theta = [-180°; +180°]$ geordneten Werte $d(\theta)$ er-



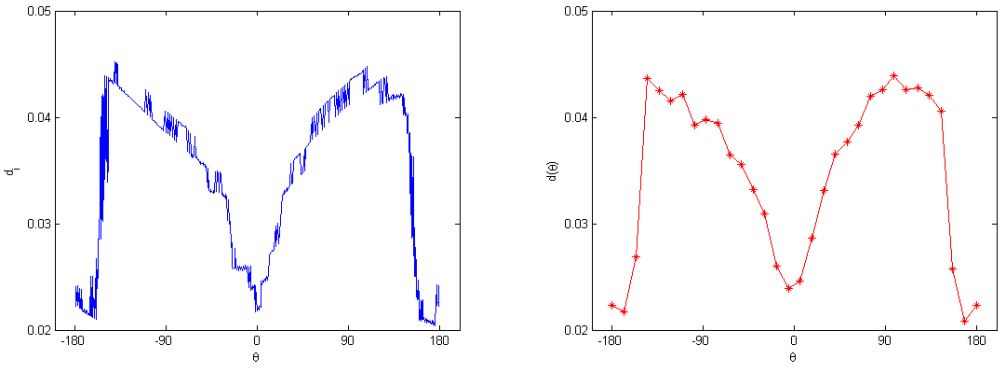
(a) (b) (c)

Abbildung 4.15.: **(a)** Seitenansicht der Ebenen E (grün) und E' (magenta). Die Ebene E' ergibt sich durch Verschieben von E in Richtung des Normalenvektors n_0. Für jeden der M diskreten Punkt von C (blau) wird ein Abstand d_i zu E' berechnet. Es gilt $i = 1, ..., M$. **(b)** Auf die Ebene E' projizierte Kurve C'. Jedem der M diskreten Punkte von C' ist ein euklidischer Abstand d_i zugeordnet. **(c)** In gleichmäßigen Abstandsintervallen $\Delta\theta$ abgetastete Kurve C'. Die den Abtastwinkeln zugeordnete Abstände $d(\theta)$ werden aus den Abständen d_i interpoliert.

geben den endgültigen Punktsignaturmerkmalsdeskriptor. Unter anderem an dieser Stelle unterscheidet sich der Ansatz von der in [Chua et al., 2000] beschriebenen Vorgehensweise, bei welcher die Punktsignaturen anschließend in eine vereinfachte Form überführt werden, um eine Personendatenbank zu erstellen. Ein weiterer Unterschied besteht darin, dass sich die Merkmalsextraktion in [Chua et al., 2000] auf starre Bereiche des Gesichts beschränkt und für jedes Voxel dieses Bereichs eine Punktsignatur ermittelt wird.

4.4.2. Experimentelle Untersuchung

Vor der Extraktion der Punktsignaturen ist es erforderlich, geeignete Werte für den Signaturradius r und das Abtastintervall $\Delta\theta$ festzulegen. Die experimentelle Auswertung gliedert sich in drei Teile. Im ersten Teil wird der Einfluss des Signaturradius auf die Klassifikationsergebnisse untersucht. Dazu wurden, bei konstantem Abtastwinkel $\Delta\theta = 11{,}6°$, acht verschiedene Radien des Intervalls [4cm; 7,5cm] getestet. Im zweiten Teil wird die Analyse um weitere Abtastintervalle $\Delta\theta \in \{5{,}7°, 24°\}$ erweitert. Im letzten Teil werden mehrere Punktsignaturen unterschiedlicher Radien kombiniert, um bei der Merkmalsextraktion einen größeren Bereich des Gesichts abzudecken. Der Ablauf der Experimente entspricht dem, im Unterkapitel 4.2 vorgestellten, allgemeinen



(a) (b)

Abbildung 4.16.: **(a)** Abstände d_i der einzelnen Punkte von C', abgetragen auf der Polarkoordinate θ. Die gezeigten Beispieldaten umfassen 560 diskrete Punkte. Aus Gründen der Übersicht wird zur Visualisierung jedoch eine kontinuierliche Linie verwendet. **(b)** In gleichmäßigen Intervallen $\Delta\theta$ abgetastete Funktion $d(\theta)$. Die Abstände der Abtastwinkel wurden mittels linearer Interpolation aus d_i gewonnen. Die von $\theta = [-180°, +180°]$ geordneten Werte $d(\theta)$ ergeben den Punktsignaturmerkmalsdeskriptor.

Testszenario.

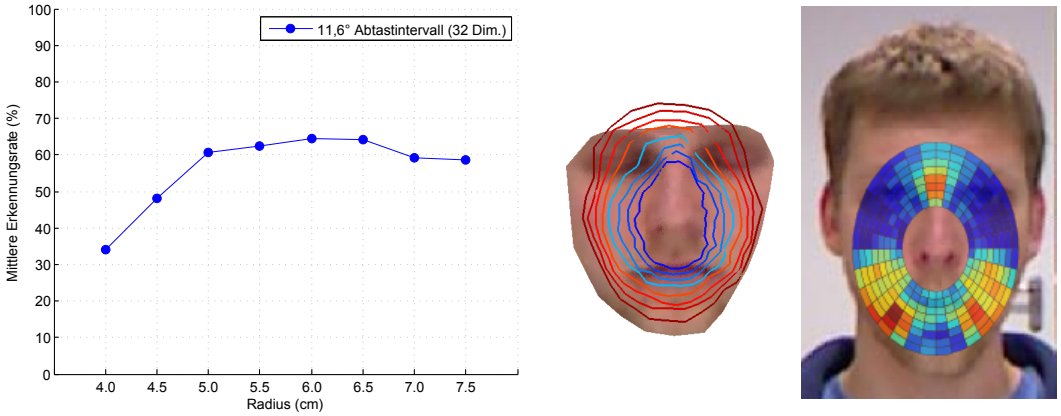
Variation des Radius r

Über den Referenzpunkt und den Radius der Punktsignatur lässt sich die Position der Extraktionsregion steuern. Auf diese Weise kann die Merkmalsextraktion wahlweise auf starre oder, wie in dieser Arbeit, auf dynamische Bereiche des Gesichts verlegt werden. Für die experimentelle Auswertung wurden in Abständen von 0,5cm insgesamt acht Radien aus dem Intervall [4 cm; 7,5cm] ausgewählt. In den Abbildungen 4.17a bis 4.17c sind beispielhaft Punktsignaturverläufe mit drei unterschiedlichen Radien zu sehen. Die Ergebnisse der Experimente zeigen, dass ein Zusammenhang zwischen dem gewählten Radius und der erzielten mittleren Erkennungsrate besteht. Die mittleren Erkennungsraten liegen zwischen 34,09 % und 64,57 %, wobei die besten Ergebnisse für $r = 6$ cm erreicht wurden. Im Schaubild 4.18a sind die Ergebnisse in Abhängigkeit vom Radius abgetragen.

Jede der acht Punktsignaturen besteht nach der Abtastung und Interpolation aus 32 diskreten Einträgen. Zur besseren Anschaulichkeit sind die Punktsignaturen in der Abbildung 4.18b durch kontinuierliche Verläufe visualisiert. Insgesamt ergeben sich $8 \cdot 32 = 256$ Merkmalsvariablen. Für jede Merkmalsvariable wurde der Zusammenhang zur Übungsklassenzielvariable mittels MI geschätzt. Die Werte wurden über alle elf

(a) (b) (c)

Abbildung 4.17.: Punktsignaturverläufe mit unterschiedlichen Radien. Das Überschreiten des Gesichtsbereiches ist visualisierungsbedingt, bei der Extraktion werden auch Informationen aus der Stirnoberfläche gewonnen. Die Größe des Radius r entspricht **(a)** 4cm **(b)** 6,5cm **(c)** 7,5cm



(a) (b) (c)

Abbildung 4.18.: **(a)** Mittlere Erkennungsraten für die Klassifikation mit Punktsignaturen verschiedener Radien. Von links nach rechts ergeben sich im Detail folgende Erkennungsraten: 34,09 %, 48,21 %, 60,65 %, 62,31 %, 64,57 %, 64,19 %, 59,17 %, 58,66 %. **(b)** Punktsignaturen der acht Radien von 4cm (dunkelblau) bis 7, 5cm (dunkelrot). **(c)** Visualisierung der Mutual-Information. Das Farbschema reicht von dunkelrot (höchste MI) bis dunkelblau (niedrigste MI). Jede der acht konzentrischen Ellipsen symbolisiert eine Punktsignatur und ist in 32 Segmente unterteilt, die die einzelnen Signaturdimensionen repräsentieren. Die MI wurde aus den Bildern aller elf Personen ermittelt. Das darunterliegende Farbbild dient nur der Visualisierung. Da die Form der Punktsignaturen vom Gesichtsausdruck abhängig ist und von der elliptischen Form abweicht, ist die räumliche Zuordnung nur als grobe Orientierung zu verstehen.

Personen des Datensatzes gemittelt. Die Ergebnisse sind in der Abbildung 4.18c visualisiert und ermöglichen eine ungefähre räumliche Zuordnung der Mutual-Information (siehe dazu Anmerkungen in der Bildunterschrift). Es zeigt sich, dass die aus Wangen- und Mundregion extrahierten Merkmalsdeskriptoren einen Zusammenhang zur ausgeführten Übung aufweisen. Für die starren seitlichen Augenregionen wurde hingegen nur eine sehr geringe MI ermittelt. Die hohe MI im Bereich der starren Nasenwurzel erscheint ungewöhnlich, lässt sich jedoch dadurch erklären, dass sich die vertikale Position einer Punktsignatur in Abhängigkeit von der ausgeführten Übung leicht ändert. So hebt sich beispielsweise beim Aufblasen der Wangen auch die Nasenspitze an, was die Punktsignatur insgesamt nach oben schiebt. Ein Vergleich der Punktsignaturverläufe aller Radien zeigt, dass die Punktsignaturen der mittleren Radien in einem Grenzbereich verlaufen. Je nach ausgewählter Übung verlaufen sie über den Nasenrücken oder über den unteren Bereich der Stirn.

Auf Basis der Konfusionsmatrix in der Abbildung 4.19 lassen sich, hier für die Experimentkonfiguration mit $r = 6$cm, die Vertauschungen zwischen den einzelnen Klassen untersuchen. Die besten, über alle Testpersonen gemittelten, übungsspezifischen Erkennungsraten werden für die Übungen *Wangen*, *WangeLi* und *WangeRe* erzielt. Wiederholte Vertauschungen treten hingegen zwischen den Übungen *Kuss*, *OForm* und *UForm* auf, welchen das Vorstülpen der Lippen gemein ist. Auch die Übungen *Breit* und *IForm* weisen erhöhte Vertauschungsraten auf. Sie ähneln sich hinsichtlich der Mundwinkelverschiebung, unterscheiden sich jedoch bei der Lippenöffnung. Achsengespiegelte Wangenübungen, wie *BoxenLi* und *BoxenRe*, sowie *WangeLi* und *WangeRe*, werden nicht verwechselt. Hier zeigt sich ein Unterschied zu den Distanzmerkmalen, die keine Informationen aus den landmarkenarmen Regionen beinhalten.

Variation des Abtastintervalls $\Delta\theta$

Eine Punktsignatur besteht, wie bereits erwähnt, aus diskreten 3D-Punkten, die sich zu einer dreidimensionalen Kurve verbinden lassen. Über die Punktanzahl kann die Detailliertheit der Kurvendarstellung gesteuert werden. Um den Einfluss der Kurvenauflösung auf die Übungsklassifikation zu untersuchen, wurden drei verschiedene Abtastintervalle $\Delta\theta \in \{24°, 11{,}6°, 5{,}7°\}$ ausgewählt und getestet. Sie ergeben Merkmalsdeskriptoren mit 16, 32 und 64 Dimensionen. Im Schaubild 4.20 ist ersichtlich, dass eine Erhöhung der Kurvenauflösung tendenziell zu einer Verbesserung der mittleren Erkennungsrate führt. Da mit dem Zuwachs der Merkmalsdimensionen ab einem gewissen Punkt jedoch auch die Gefahr des Übertrainings zunimmt, kann die Integration eines Merkmalsreduktionsverfahren sinnvoll sein.

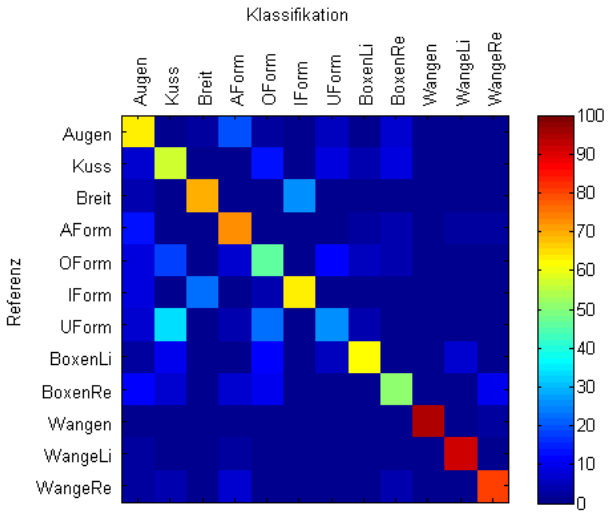


Abbildung 4.19.: Konfusionsmatrix für $r = 6\text{cm}$. Die mittlere Erkennungsrate der Diagonaleinträge beträgt 64,57 %.

Konkatenierung mehrerer Punktsignaturen

Durch die Extraktion mehrerer Punktsignaturen mit unterschiedlichen Radien r, kann die Informationsextraktion auf einen größeren Teil des Gesichts ausgedehnt werden (vgl. Abb. 4.17 und 4.18b). Die Konkatenierung aller acht Punktsignaturen zu einem Merkmalsvektor resultiert in einer Verbesserung der mittleren Erkennungsraten auf bis zu 75,40 % (für $\Delta\theta = 5{,}7°$). Eine Gegenüberstellung der besten Einzelergebnisse mit den Resultaten der Konkatenierung findet sich im Schaubild 4.21. Die Konkatenierung mehrerer Punktsignaturen hat jedoch auch eine Erhöhung der Dimensionsanzahl zur Folge, weshalb die zusätzliche Integration einer Merkmalsreduktion sinnvoll sein kann.

4.4.3. Fazit

Die Extraktion von einzelnen Punktsignaturen ermöglicht mittlere Erkennungsraten von bis zu 65,74 %. Eine Konkatenierung von Deskriptoren unterschiedlicher Radien verbessert die MER auf bis zu 75,40 %, da die Signaturverläufe größere Bereiche des Gesichts abdecken. Anders als Distanzmerkmale enthalten Punktsignaturen auch Informationen über Oberflächenverläufe in landmarkenarmen, homogenen Regionen des Gesichts. Dies führt zu einer deutlich besseren Unterscheidung der achsengespiegelten Wangenübungen.

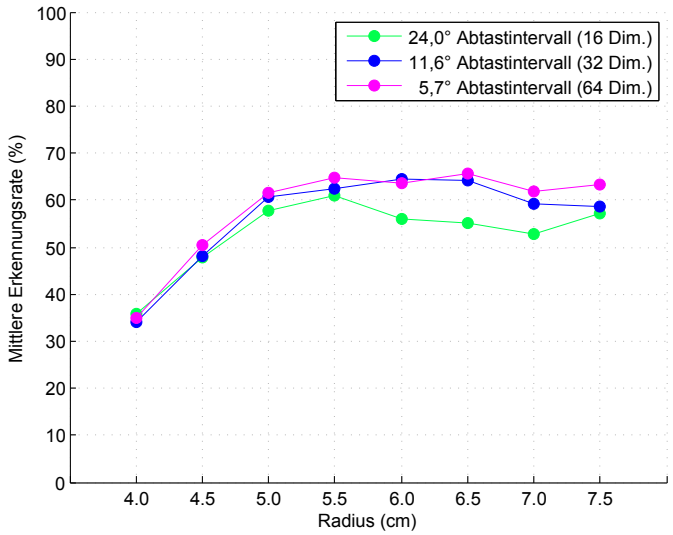


Abbildung 4.20.: Vergleich der mittleren Erkennungsraten für verschiedene Abtastintervalle $\Delta\theta$.

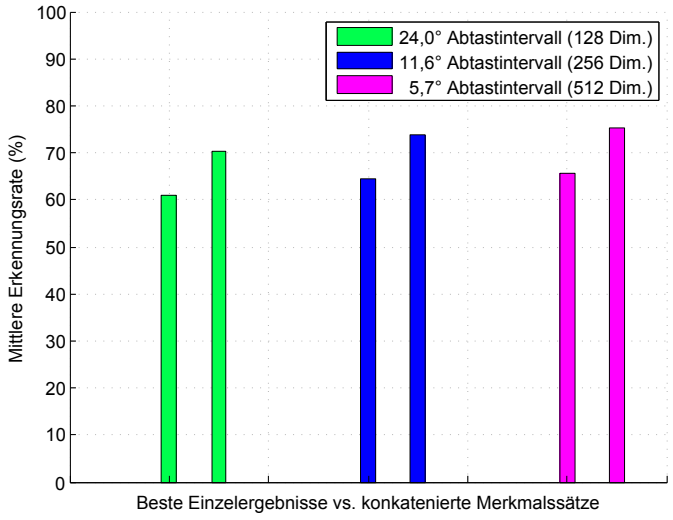


Abbildung 4.21.: Gegenüberstellung der Ergebnisse für die konkatenierten Punktsignaturen (rechts) mit den besten Einzelergebnissen (links). Für Signaturen mit dem Abtastintervall $\Delta\theta = 24°$ ergibt sich eine verbesserte MER von 70,25 % im Vergleich zu 60,93 % für $r = 5{,}5$cm. Die Merkmalsgruppen mit $\Delta\theta = 11{,}6°$ ergeben, im Vergleich zum Einzelergebnis 64,57 % ($r = 6$cm), eine MER von 73,89 %. Die insgesamt beste MER von 75,40 % wird für $\Delta\theta = 5{,}7°$ erzielt (Einzelergebnis 65,74 %, $r = 6{,}5$cm).

4.5. Histogramme orientierter Normalenvektoren

Als HON-Merkmale (*engl.* Histograms of Oriented Normals) werden Histogramme aus orientierten Normalenvektoren bezeichnet. Da die Merkmalsextraktion patchbasiert erfolgt, ist keine vorgeschaltete Landmarkenlokalisierung notwendig. Die mit Hilfe der Gesichtsdetektion gewonnene Bounding Box wird dabei in r Extraktionsregionen unterteilt, für welche jeweils ein Histogramm ermittelt wird. Das Extraktionsverfahren wurde im Rahmen dieser Arbeit implementiert und orientiert sich an [S. Tang et al., 2013], [Berretti et al., 2011] und [Lemaire et al., 2013].

Der Rest dieses Unterkapitels gliedert sich in einen Theorieteil zur HON-Adaption dieser Arbeit (Kap. 4.5.1) und in eine detaillierte, experimentelle Evaluation (Kap. 4.5.2). Abschließend werden im Unterabschnitt 4.5.3 die wichtigsten Aspekte zusammengefasst.

4.5.1. Theorie und Extraktion

Die Beschaffenheit einer Oberfläche kann durch ihre Normalenvektoren beschrieben werden. Eine Veränderung der Oberfläche hat Einfluss auf die Ausrichtung der Normalenvektoren, wie Abbildung 4.22a zeigt. Die Lage eines Normalenvektors $\mathbf{N}$ im dreidimensionalen Raum wird, in Anlehnung an die sphärischen Koordinaten, durch die Winkel Azimut φ und Elevation θ beschrieben. Diese sind so definiert, dass der Azimutwinkel $\varphi = [-180°, +180°]$ die Richtung innerhalb der xy-Ebene kennzeichnet und der Elevationswinkel $\theta = [-90°, +90°]$ auf der xy-Ebene aufsetzt[9]. Die Abbildung 4.22b visualisiert beide Winkel in einem dreidimensionalen Koordinatensystem. Zur Anpassung an den Datensatz, sowie einer intuitiveren Interpretation, weicht die Notation innerhalb dieser Arbeit in geringem Maße von dieser Darstellung ab. Details dazu finden sich an späterer Stelle.

Um die Winkel φ und θ für verschiedene Gesichter valide vergleichen zu können, ist ein Referenzkoordinatensystem notwendig. In diesem werden alle Punktwolken einheitlich ausgerichtet. Ein zufällig ausgewähltes Gesicht des Datensets dient dabei, sowohl während des Trainings- als auch des Testprozesses, als Referenzmodell. Mit Hilfe des Iterative-Closest-Point-Algorithmus (ICP) werden alle Punktwolken so verschoben und rotiert, dass ihr Abstand zur Referenzpunktwolke minimiert wird (Informationen zum ICP-Algorithmus siehe Seite 66).

Da die einzelnen Gesichter des Datensatzes im Prinzip nur aus Ansammlungen von 3D-Punkten bestehen, muss vor der Bestimmung der Normalenvektoren eine Ober-

[9] Festlegung nach `http://de.mathworks.com/help/matlab/ref/cart2sph.html`, letzter Zugriff: 24.03.2015

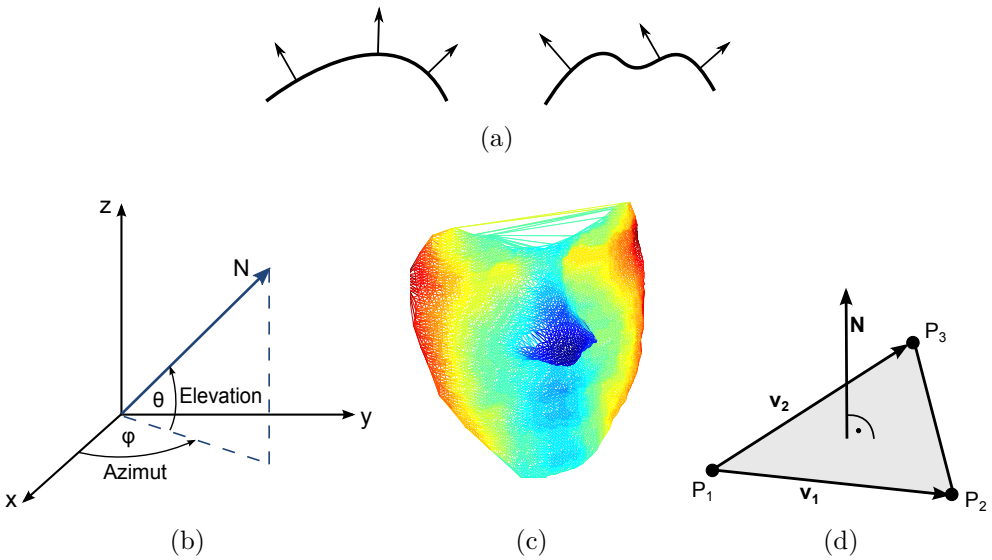


Abbildung 4.22.: **(a)** Ausrichtung der Normalenvektoren in Abhängigkeit von der Oberflächenbeschaffenheit. **(b)** Definition von Azimut- und Elevationswinkel im dreidimensionalen Raum. **(c)** Ein nach der Delaunay-Triangulation aus einem Dreiecksnetz bestehendes Beispielgesicht. **(d)** Aus zwei Seiten eines Dreiecks lässt sich mit Hilfe des Kreuzproduktes $\mathbf{v_1} \times \mathbf{v_2}$ der Normalenvektor $\mathbf{N}$ der Dreiecksfläche berechnen.

fläche aus diesen Punkten geschätzt werden. Dazu wird eine Delaunay-Triangulation angewandt, die eine durch Dreiecksflächen repräsentierte Polygonoberfläche erzeugt[10]. Ein Beispiel für eine resultierende Polygonoberfläche, bestehend aus circa 10 000 Dreiecken, findet sich in Abbildung 4.22c. Für jede Dreiecksfläche kann im nächsten Schritt mit Hilfe des Kreuzproduktes ein Normalenvektor $\mathbf{N}$ aus den Vektoren von zwei beliebigen Dreiecksseiten $\mathbf{v_1}$ und $\mathbf{v_2}$ geschätzt werden:

$$\mathbf{v_1} \times \mathbf{v_2} = \mathbf{N}. \tag{4.7}$$

Die schematische Darstellung dieser Vorgehensweise ist in Abbildung 4.22d zu sehen. Da es sich um ein Rechtssystem handelt, beeinflusst die Richtung der Vektoren $\mathbf{v_1}$ und $\mathbf{v_2}$ das Vorzeichen des Kreuzproduktes, was im Ergebnis zu uneinheitlich ausgerichteten Normalenvektoren führen kann. Um eine einheitliche Orientierung zu erreichen, werden alle $\mathbf{N}$, für die gilt $z > 0$, mit dem Faktor -1 multipliziert (siehe Abb. 4.23a). Die neu ausgerichteten Normalenvektoren sitzen nun ebenfalls auf der konvexen Außenseite der Gesichtsoberfläche auf.

Nach erfolgter Ausrichtung beginnt die Extraktion der Winkelpaare (Abb. 4.23b

[10]Für die Delaunay-Triangulation wurde die von Matlab bereitgestellte Funktion *delaunay* verwendet (siehe `http://de.mathworks.com/help/matlab/ref/delaunay.html`, letzter Zugriff: 17.03.2015).

und 4.23c). Ein Beispielgesicht mit 10 000 Polygonflächen und je einem Normalenvektor pro Fläche ergibt einen Merkmalsvektor von $2 \times 10\,000$ Dimensionen. Zur Reduktion dieser hohen Datenmenge werden die Winkel in einem bivariaten Histogramm, bestehend aus $m \times n$ Bins, zusammengefasst. Abbildung 4.23d zeigt ein Beispielhistogramm mit $m = 5$ und $n = 5$ Bins.

Ein Nachteil der Histogrammbildung besteht darin, dass die räumliche Zuordnung der Merkmale verloren geht. Da diese jedoch ebenfalls zur Erkennung der ausgeführten Übung beitragen kann, wird das Gesicht in kleinere Extraktionsregionen unterteilt. Jeder Region ist ein eigenes bivariates Histogramm zugeordnet. Die Länge des Merkmalsvektors ist sowohl abhängig von der Histogrammaufteilung als auch von der Anzahl der Extraktionsregionen und umfasst $m \times n \times r$ Dimensionen. Im folgenden Evaluationskapitel werden verschiedene Aufteilungen des Gesichts experimentell untersucht.

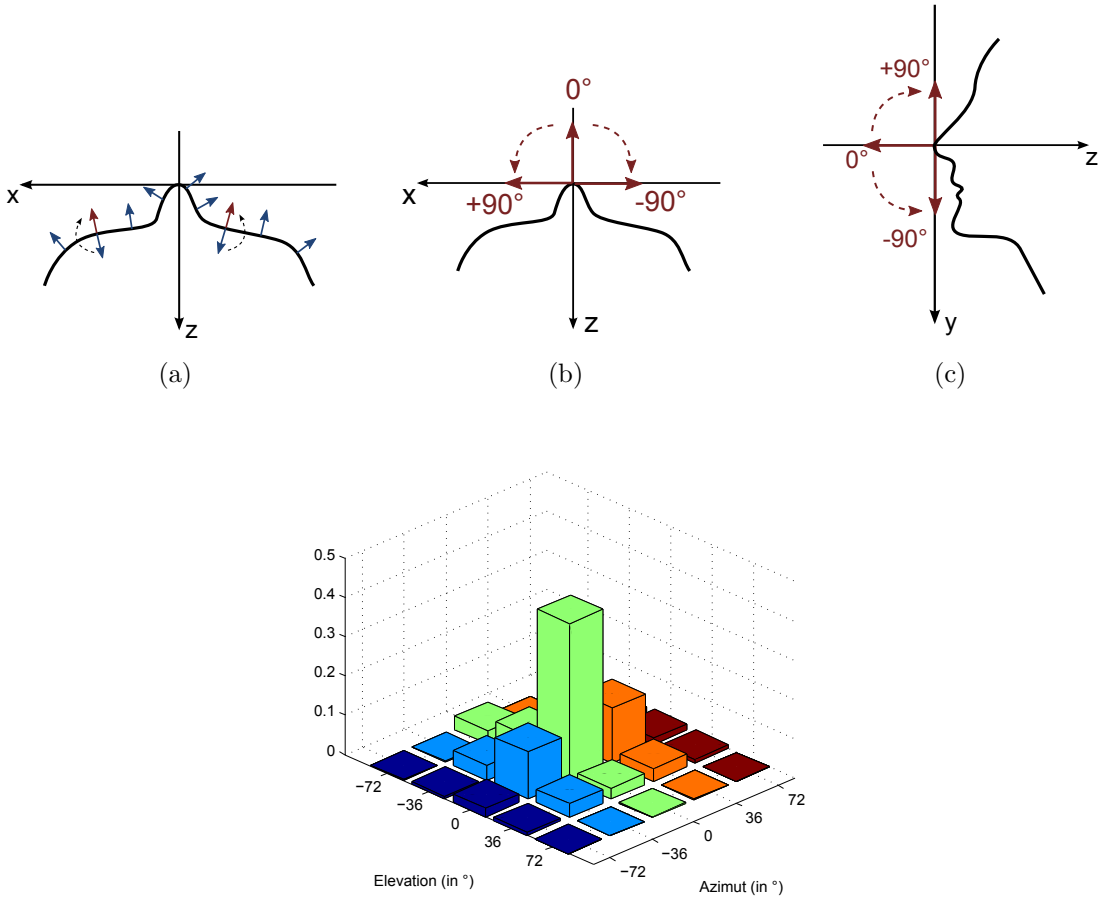
4.5.2. Experimentelle Untersuchung

Bei der Extraktion der HON-Deskriptoren sind zahlreiche Abwandlungen, beispielsweise hinsichtlich der Regionen- oder Histogrammaufteilung, möglich. Der allgemeinen Testvorgehensweise im Unterkapitel 4.2 folgend, wurden mehrere dieser Varianten experimentell untersucht. Die Dokumentation der Ergebnisse teilt sich in zwei Abschnitte auf. Der erste Abschnitt befasst sich detaillierter mit der Anzahl der Extraktionsregionen, da ihre Erhöhung den deutlichsten positiven Einfluss auf die mittlere Erkennungsrate zeigte. Der zweite Abschnitt enthält eine kurze Übersicht über weitere Modifikationen, die keinen beziehungsweise einen geringen Beitrag zur Übungserkennung ergaben, jedoch der Vollständigkeit halber Erwähnung finden sollen.

Anzahl der Extraktionsregionen

Das Zusammenfassen der extrahierten Deskriptoren in einem Histogramm führt dazu, dass der räumliche Bezug dieser Deskriptoren verloren geht. Um dessen Einfluß auf die Übungsklassifikation zu untersuchen, wurden die Tests mit je $r \in \{2, 4, 6, 8\}$ Extraktionsregionen durchgeführt. Die verschiedenen Varianten sind in Abbildung 4.24 gezeigt. Basierend auf den Mutual-Information-Schätzungen der in den vorhergehenden Unterkapiteln beschriebenen Deskriptoren, wurde für die Mundregion eine feinere und die Augenregion eine gröbere Aufteilung gewählt (vgl. Abb. 4.8 und 4.18c). Andere Aufteilungen, beispielsweise in Form eines gleichmäßigen Gitters, wären jedoch ebenfalls möglich.

Die erzielten mittleren Erkennungsraten liegen für die Klassifikation mit zwölf



(a) (b) (c)

(d)

Abbildung 4.23.: **(a)** Die schematische Zeichnung zeigt eine Draufsicht auf den Kopf mit angedeutetem Verlauf der Wangen, sowie der Nasenspitze im Ursprung des Koordinatensystems. Die über das Kreuzprodukt ermittelten Normalenvektoren (blau) weisen in unterschiedliche Richtungen. Um eine einheitliche Orientierung zu erreichen, werden alle $\mathbf{N}$, für die gilt $z > 0$, mit dem Faktor -1 multipliziert. Die neu ausgerichteten Normalenvektoren (rot) sitzen nun ebenfalls auf der konvexen Außenseite der Gesichtsoberfläche auf. **(b)** Visualisierung des Azimutwinkels, welcher für das Intervall $[-90°, +90°]$ definiert ist. Im Gegensatz zur allgemeinen Notation (siehe Abb. 4.22b) liegt der Azimut des HON-Merkmals in dieser Arbeit innerhalb der xz-Ebene. **(c)** Der Elevationswinkel setzt auf der xz-Ebene auf und ist für das Intervall $[-90°, +90°]$ definiert. **(d)** Bivariates Histogramm der Winkelpaare Azimut und Elevation. Die Achsenbeschriftungen zeigen die Werte der Binzentren an.

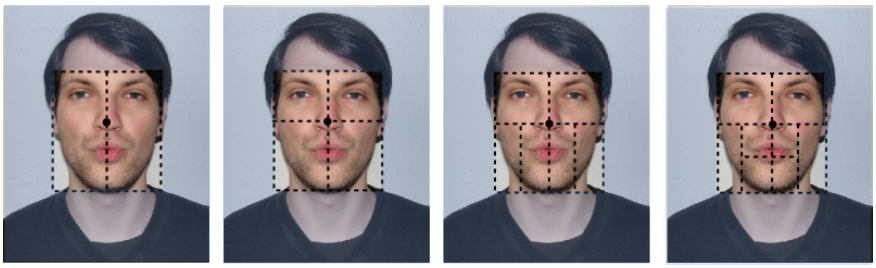
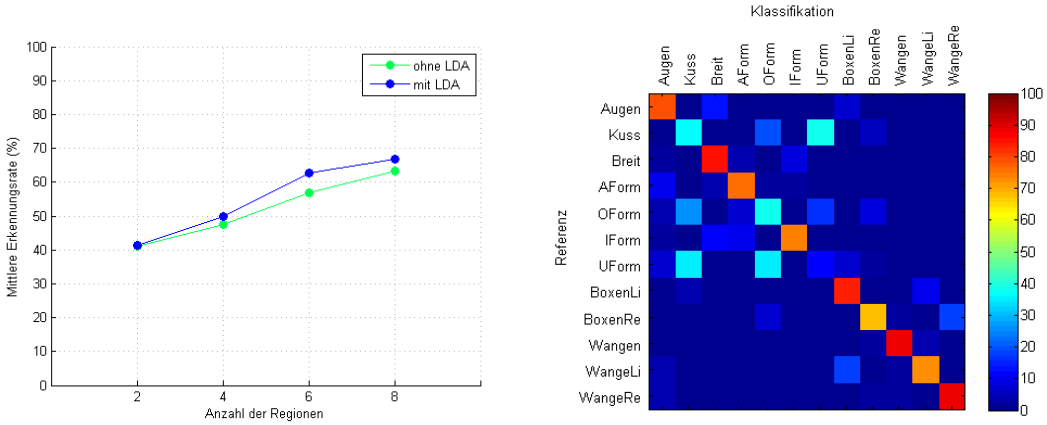


Abbildung 4.24.: Unterteilung des Gesichts in $r \in \{2, 4, 6, 8\}$ Regionen.

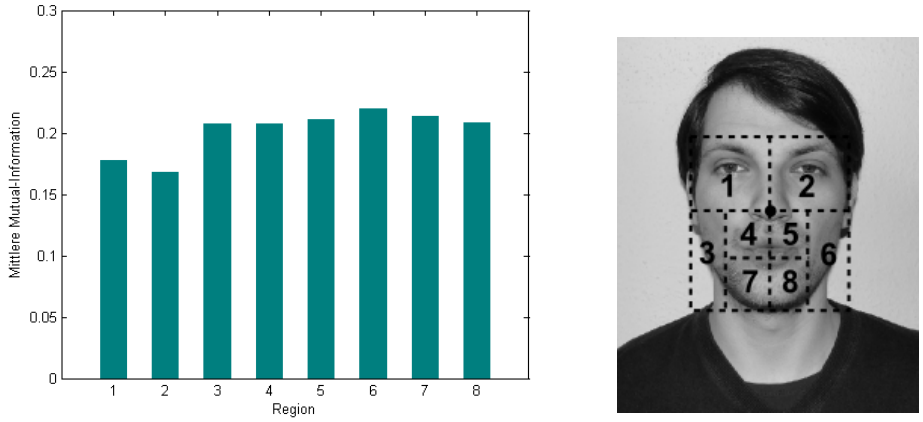


(a) (b)

Abbildung 4.25.: **(a)** Einfluss der Regionenanzahl auf die MER. **(b)** Konfusionsmatrix (Acht Extraktionsregionen, zwölf Klassen, MER: 66,79 %).

Übungsklassen zwischen 41,12 % und 63,41 %. In Abbildung 4.25a wird deutlich, dass die feinste Unterteilung des Gesichts mit $r = 8$ die höchste MER erzielt. Da die Länge des Merkmalsvektors in diesem Fall $m \times n \times r = 200$ Dimensionen umfasst, wurde zur Reduktion der Merkmalsdimensionen eine lineare Diskriminanzanalyse (LDA) vorgeschaltet. Diese führt eine Transformation des Merkmalsraumes durch und reduziert diesen dabei auf $k - 1$ Dimensionen, wobei $k = 12$ der Anzahl der Klassen entspricht. Mit Hilfe der LDA kann eine Verbesserung der Ergebnisse auf 66,79 % für $r = 8$ und zwölf Klassen erzielt werden.

Um die Vertauschungen zwischen den einzelnen Klassen zu untersuchen, ist in Abbildung 4.25b die Konfusionsmatrix für die Klassifikation mit zwölf Klassen und acht Regionen visualisiert (MER: 66,79 %). Der größte Schwachpunkt des Ansatzes zeigt sich bei der Unterscheidung der Übungen mit gespitzten Lippen (*Kuss*, *OForm*, *UForm*). Hier rangieren die übungsspezifischen Erkennungsraten zwischen 11,69 % und 37,96 %. Dem gegenüber stehen allein 38,75 % der *Kuss*-Observationen, die fälschlicherweise als *UForm* klassifiziert werden. Für die restlichen Übungsklassen zeigen sich deutliche



(a) (b)

Abbildung 4.26.: **(a)**–**(b)** Über alle 25 Histogrammbins gemittelte MI der acht Regionen.

bessere MERs (67,97 % bis 88,74 %).

Um den Zusammenhang zwischen den extrahierten Merkmalsdeskriptoren und den ausgeführten Übungen zu analysieren, wurde für jede der 200 Merkmalsdimensionen die Mutual-Information zur Zielvariable geschätzt. Die 200 Ergebniswerte wurden anschließend zu acht regionenspezifischen Mittelwerten zusammengefasst. Die Abbildung 4.26a zeigt diese in Form eines Säulendiagramms. Analog zu den Ergebnissen der vorhergehenden Merkmalsextraktionen, weisen die Merkmale der Mund-Wangen-Regionen höhere MI-Werte als die Augenregionen auf (vgl. Abb. 4.8 und 4.18c).

Weitere Variationen

Neben der Variation der Regionenanzahl wurden drei weitere Modifikationsmöglichkeiten untersucht. Diese erzielten keine zufriedenstellenden Ergebnisse, sollen der Vollständigkeit halber aber dennoch Erwähnung finden.

- Die Erhöhung der Binanzahl des bivariaten Histogramms auf 7×7 führte zu einer Abnahme der mittleren Erkennungsrate von 63,41 % auf 58,57 %. Ein möglicher Grund hierfür könnte sein, dass der Auflösungszuwachs neben den zentralen Bins auch Randbins betrifft. Diesen werden jedoch nur wenige Normalenvektoren zugeordnet (siehe dazu Abb. 4.23d). Bei einem Zuwachs der Dimensionsanzahl von 200 auf 392 erhöht sich somit auch die Anzahl der Dimensionen mit geringer Relevanz. Um diese Vermutung zu überprüfen, könnte zum Vergleich ein bivariates Histogramm mit nicht-gleichmäßiger Binauflösung getestet werden.

- Die an [Lowe, 2004] angelehnte, binübergreifende Wichtung für Samples in

Bingrenzen-Nähe führte zu einer leichten Verschlechterung von 63,41 % auf 63,01 %.

- Eine Wichtung der Histogrammbins mit dem Raumwinkel, analog zu [Scovanner et al., 2007], wurde ebenfalls untersucht und führte zu einer Verschlechterung auf 51,91 %. Dies könnte dadurch bedingt sein, dass insbesondere Winkel in den Randbereichen stärker gewichtet werden, in welche nach Abbildung 4.23d jedoch nur wenige Normalenvektoren fallen.

Eine tiefergehende, experimentelle Evaluation der aufgezählten Modifikationen und ihrer Schwachpunkte übersteigt den Fokus dieser Arbeit, kann als Ansatzpunkt für nachfolgende Arbeiten jedoch sinnvoll sein.

4.5.3. Fazit

Die Klassifikation auf Basis der extrahierten HON-Merkmalsdeskriptoren erzielte vergleichbare Ergebnisse wie die Klassifikation mit Punktsignaturen und DW-Merkmalen (MER: 63,41 %, 66,79 % (LDA)). Die HON-Extraktion ist patchbasiert und benötigt im Grunde keine vorgeschaltete Landmarkenlokalisierung. Im Rahmen der Vorverarbeitung ist die Kenntnis einzelner Landmarken dennoch von Vorteil, da die Winkelbestimmung eine möglichst einheitliche Registrierung aller Gesichtsmodelle in einem Referenzkoordinatensystem erfordert.

4.6. Krümmungsmerkmale

Die Krümmungsanalyse umfasst zahlreiche Methoden und Vorgehensweisen zur Beschreibung von Oberflächeneigenschaften. In den folgenden vier Abschnitten wird eine Auswahl dieser Methoden vorgestellt und evaluiert. Für weiterführende Informationen sei auf [Besl und R. C. Jain, 1986], [Flynn und A. K. Jain, 1989], [Chen und Schmitt, 1992], [Barth et al., 1993] und [Cantzler und Fisher, 2001] verwiesen. Der erste Abschnitt dieses Unterkapitels enthält eine theoretische Übersicht zu ausgewählten Krümmungsdeskriptoren. Im Anschluss wird die Krümmungsschätzung für reale Tiefendaten beschrieben. Der dritte Abschnitt dokumentiert die experimentelle Evaluation der implementierten Verfahren. Im abschließenden Fazit werden die wichtigsten Aspekte zusammengefasst.

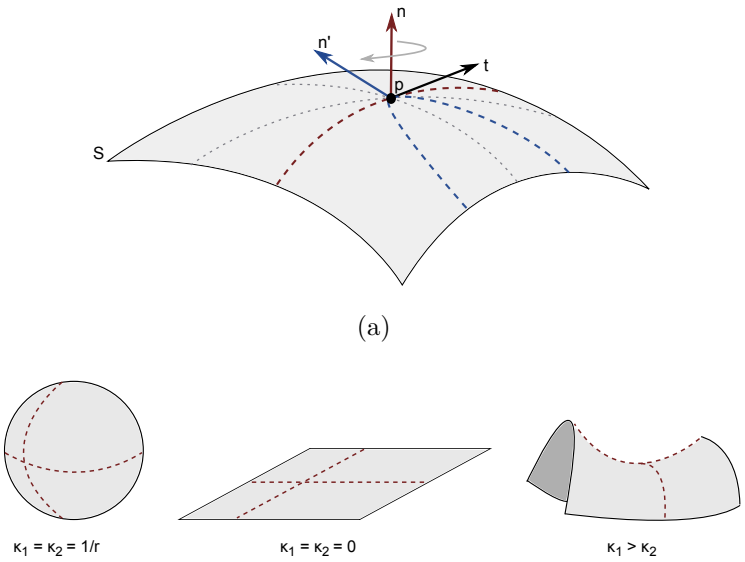


Abbildung 4.27.: **(a)** Oberfläche mit vier verschiedenen Krümmungsverläufen durch den Punkt **p**. Jeder Verlauf besitzt in **p** einen eigenen charakteristischen Krümmungswert κ. **(b)** Die Relation der Hauptkrümmungen zueinander am Beispiel einer Kugel, einer planaren Ebene und einer Sattelfläche.

4.6.1. Grundlagen und Konzept der Krümmungsextraktion

In dieser Arbeit werden zur Analyse einer Oberfläche die mittlere Krümmung H und die Gauß'sche Krümmung K extrahiert und anhand der sogenannten HK-Klassifikation in intuitiv interpretierbare Krümmungskategorien unterteilt. Die Grundlage für diese (und weitere) Deskriptoren bilden die beiden Hauptkrümmungen κ_1 und κ_2, die im Folgenden zuerst beschrieben werden.

Als Ausgangspunkt der Betrachtungen dient die in Abbildung 4.27a gezeigte Oberfläche S. Sie enthält vier verschiedene Beispielkurven, die allesamt den Punkt **p** passieren. Die blaue und die rote Kurve verfügen über denselben Tangentenvektor **t**, unterscheiden sich jedoch in der Richtung ihrer Normalenvektoren **n** und **n′**. Alle folgenden Ausführungen beschränken sich auf Kurvenverläufe, deren Normalenvektoren, analog zu **n**, senkrecht auf einer Oberfläche stehen. Durch Rotation der roten Kurve um die Achse von **n** entstehen weitere Kurven, die jeweils einen eigenen Krümmungswert κ in **p** besitzen [Chen und Schmitt, 1992]. Die gestrichelten, grauen Linien zeigen zwei Beispiele dieser Kurvenverläufe.

Die beiden Kurven mit dem maximalen und minimalen Krümmungswert in **p** verlaufen stets senkrecht zueinander. Man bezeichnet diese Extremwerte auch als die *Hauptkrümmungen* κ_1 und κ_2 (*engl.* principal curvatures). Während im Allgemeinen

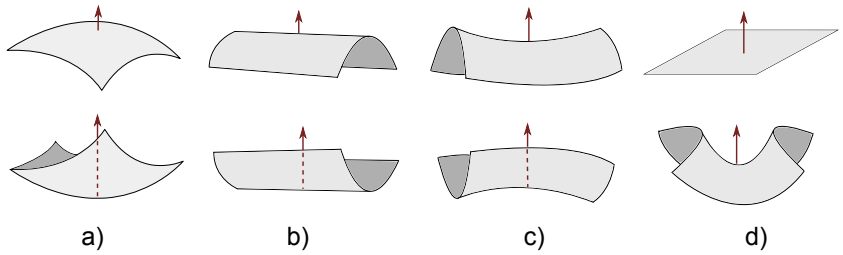


Abbildung 4.28.: Die acht möglichen Klassen, die sich aus der HK-Klassifikation nach [Colombo et al., 2006] ergeben. Der rote Pfeil zeigt die Richtung der z-Achse an. **(a)** Elliptisch konkav (Bsp. oben) bzw. elliptisch konvex (Bsp. unten) **(b)** Zylindrisch konkav bzw. zylindrisch konvex **(c)** Hyperbolisch konkav bzw. hyperbolisch konvex **(d)** Planar bzw. hyperbolisch symmetrisch.

die Relation $\kappa_1 > \kappa_2$ Gültigkeit hat, können beide in Spezialfällen auch identische Werte annehmen, wie Abbildung 4.27b zeigt. So ist für eine Kugeloberfläche im Punkt $\mathbf{p}$ die Beziehung $\kappa_1 = \kappa_2 = \frac{1}{r}$ definiert, wobei r dem Radius der Kugel entspricht. Bei einer planaren Ebene gilt $\kappa_1 = \kappa_2 = 0$. Punkte mit identischen Hauptkrümmungen werden auch als umbilische Punkte bezeichnet [Besl und R. C. Jain, 1986].

Aus den Hauptkrümmungen lassen sich weitere Krümmungsdeskriptoren ableiten. Für die *Gauß'sche Krümmung* K (*engl.* Gaussian curvature) und die *mittlere Krümmung* H (*engl.* mean curvature) gilt:

$$K = \kappa_1 \cdot \kappa_2 \tag{4.8}$$

und

$$H = \frac{\kappa_1 + \kappa_2}{2}. \tag{4.9}$$

Anschließend kann auf Basis der Skalare H und K eine sogenannte *HK-Klassifikation* durchgeführt werden ([Besl und R. C. Jain, 1986], [Colombo et al., 2006]). Diese ordnet jedem Element der Punktwolke eine der in der Abbildung 4.28 gezeigten kategorischen Klassen zu. Die exakte Klassifikationsvorschrift findet sich in Tabelle 4.4.

Aus der Tabelle wird ersichtlich, dass das Vorzeichen der mittleren Krümmung H eine Aussage über die Orientierung der Oberfläche in Relation zu einer Referenzrichtung (hier der z-Achse) erlaubt. Bei der Konkavität oder Konvexität handelt es sich somit um eine extrinsische Eigenschaft, welche eine Oberfläche in ein umliegendes (Koordinaten-)System einordnet [Besl und R. C. Jain, 1986]. Die Gauß'sche Krümmung K hingegen bestimmt den Oberflächentyp, wie z.B. Sattel oder Gipfel, und

Tabelle 4.4.: HK-Klassifikation nach [Colombo et al., 2006], wie sie auch in dieser Arbeit verwendet wird. Im Vergleich zur Vorgehensweise von [Besl und R. C. Jain, 1986] ist das Vorzeichen von H, und somit die Zuordnung von Konvexität und Konkavität vertauscht.

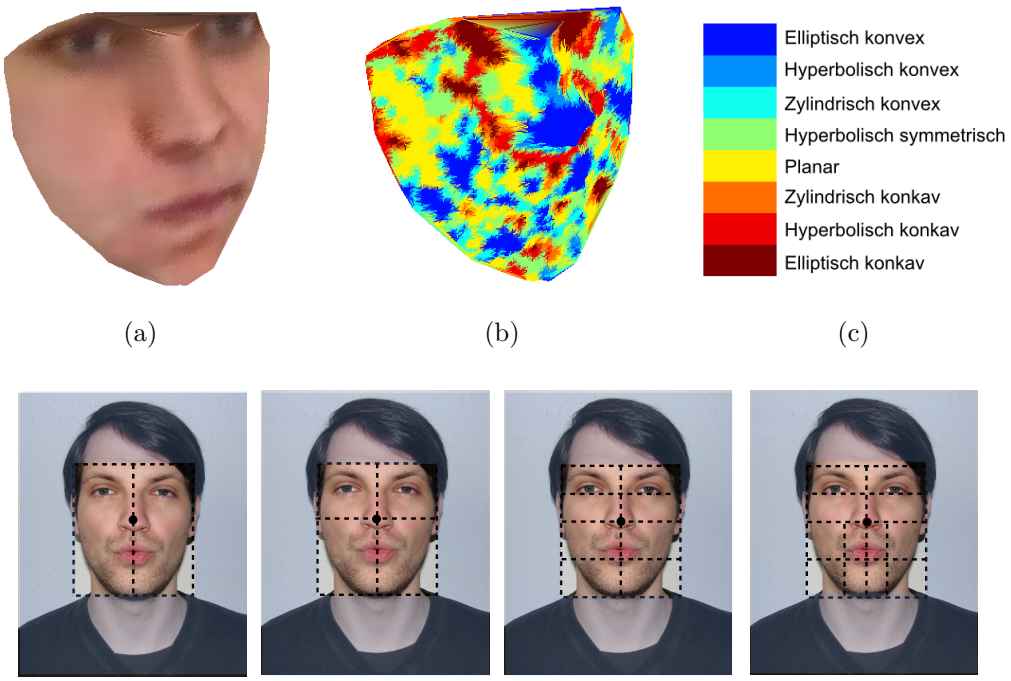
	$K < 0$	$K = 0$	$K > 0$
$H < 0$	hyperbolisch konkav	zylindrisch konkav	elliptisch konkav
$H = 0$	hyperbolisch symmetrisch	planar	nicht möglich
$H > 0$	hyperbolisch konvex	zylindrisch konvex	elliptisch konvex

ist eine intrinsische Eigenschaft, da sie unabhängig von der Lage der Oberfläche im umliegenden System ist.

Die Krümmungsberechnung basiert auf den Ableitungen erster und zweiter Ordnung und ist dadurch relativ fehleranfällig gegenüber einer verrauschten Datenbasis [Besl und R. C. Jain, 1986]. Das Zusammenfassen der Daten in kategorische Krümmungsklassen ermöglicht daher nicht nur eine intuitivere Interpretation der Ergebnisse, sondern reduziert den Einfluss des Datenrauschens. Dies legen auch eigene Experimente nahe. Eine tiefergehende Untersuchung wäre an dieser Stelle notwendig, sprengt jedoch den Rahmen dieser Arbeit.

Nachdem jedem der n Elemente einer Punktwolke eine Krümmungskategorie zugeordnet wurde (siehe Abb. 4.29a bis 4.29c), würde ohne weitere Nachverarbeitungsschritte ein n-dimensionaler Merkmalsvektor vorliegen. Da die zur Merkmalsextraktion vorbereiteten Punktwolken in dieser Arbeit zwischen 2000 und 3000 3D-Punkte umfassen, entstünde so ein sehr hochdimensionaler Merkmalsraum. Bei einem vergleichsweise kleinen Datensatz begünstigt dies eine Überanpassung des Klassifikatormodells an die Trainingsdaten. Zusätzlich unterscheidet sich die Elementanzahl zwischen den einzelnen Punktwolken. Daher werden die n-dimensionalen Merkmalsvektoren, analog zur Vorgehensweise in [J. Wang et al., 2006], in Histogrammen zusammengefasst. Da jede Krümmungskategorie ein eigenes Bin erhält, ergibt dies einen Merkmalsvektor mit acht Einträgen.

Der Nachteil dieser Vorgehensweise besteht jedoch darin, dass Information über die räumliche Zuordnung der Krümmungen verloren geht. Um dem entgegenzuwirken, wird das Gesicht in r Extraktionsregionen, sogenannte Patches, unterteilt und für jedes Patch ein eigenes Histogramm ermittelt [J. Wang et al., 2006]. Daraus ergibt sich ein $8 \times r$ dimensionaler Merkmalsvektor. Um achsengespiegelte Übungen, wie beispielsweise *WangeRe* und *WangeLi*, unterscheiden zu können sind mindestens $r = 2$ vertikal unterteilte Regionen notwendig. Die in der Abbildung 4.29d gezeigten Regionenaufteilungen werden im Experimentalteil dieses Unterkapitels evaluiert. Jedes Histogramm wird mit der Anzahl seiner Einträge normalisiert, um eine Aussage



(a) (b) (c)

(d)

Abbildung 4.29.: **(a)** Punktwolke für die Beispielübung *WangeRe*. **(b)** Falschfarbendarstellung einer Punktwolke, in welcher jedes Pixel entsprechend seiner ermittelten Krümmungsklasse eingefärbt ist. **(c)** Farblegende der acht Krümmungsklassen. **(d)** Aufteilung des Gesichts in $r \in \{2, 4, 8, 12\}$ Extraktionsregionen.

über die prozentuale Häufigkeit der Klassen zu ermöglichen.

4.6.2. Implementierung und Krümmungsschätzung auf realen Daten

Um die Vergleichbarkeit der auf verschiedenen Punktwolken geschätzten Krümmungswerte zu garantieren, sind Vorverarbeitungsschritte notwendig. Die Einbettung der Vorverarbeitungsschritte in den Ablauf der Merkmalsextraktion ist in Abbildung 4.30a gezeigt.

Um die Lage aller Punktwolken im Koordinatensystem anzugleichen, werden diese mit Hilfe des ICP-Algorithmus an einer Referenzpunktwolke ausgerichtet (Informationen zum ICP-Algorithmus siehe Seite 66). Die Referenzpunktwolke wurde aus dem Datensatz zufällig ausgewählt und bleibt über alle Experimente hinweg unverändert.

Dem Namen entsprechend, handelt es sich bei einer Punktwolke nicht um eine Oberfläche im eigentlichen Sinne, sondern um eine Ansammlung von n diskreten Punkten

$\mathbf{p}_i = (X_i\ Y_i\ Z_i)^\mathsf{T}$, mit $i = 1, ..., n$, im dreidimensionalen Raum. Für die Vergleichbarkeit der Krümmungswerte ist eine ähnliche räumliche Auflösung der Punktwolken von Bedeutung. Aus diesem Grund werden die Punktwolken neu gesampelt. Dazu muss der Verlauf der Oberfläche bekannt sein. Da dieser aus den Punktkoordinaten allein nicht hervorgeht und auch die euklidische Distanz keine Aussagekraft über die Nachbarschaft von Punkten auf der Oberfläche S enthält, dient eine Parameterebene E als Behelf. Diese liegt in unveränderlicher Relation zu S und bildet ein gleichabständiges Referenzgitter mit den Koordinaten (u, v), wie die Abbildung 4.30b zeigt. Die xy-Ebene eignet sich als Referenzebene und wird entsprechend des (u, v)-Gitters neu abgetastet (siehe Abb. 4.30c). Die u- und v-Achsen sind jedoch, anders als beim 2.5D-Bild, nicht in Pixel, sondern in Meter unterteilt und der Samplingabstand beträgt 2mm. Die vollständigen Koordinaten der neuabgetasteten 3D-Punkte $\mathbf{s}(u, v) = (u\ v\ z(u, v))^\mathsf{T}$ ergeben sich dann durch Interpolation der $z(u, v)$-Werte auf Basis der alten Tiefenwerte Z_i. Die neue Punktwolke setzt sich aus n' 3D-Punkten zusammen.

Die Darstellung einer Oberfläche über eine Parameterebene in Kombination mit Abstandswerten wird auch als Monge-Patch bezeichnet. Sie vereinfacht die Krümmungsschätzung wesentlich. Näheres dazu findet sich in [Besl und R. C. Jain, 1986].

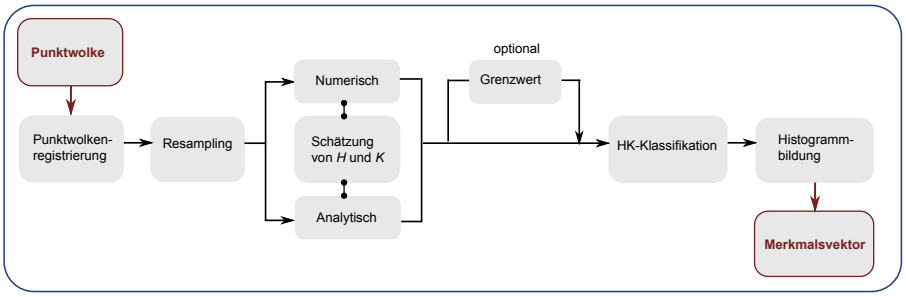
Nach der Vorbereitung der Daten erfolgt die Schätzung der Krümmungen H und K im Punkt $\mathbf{s}(u, v)$ unter Einbezug der Nachbarpunkte. Dafür existieren zahlreiche Vorgehensweisen, von denen in dieser Arbeit eine numerische und eine analytische Variante implementiert wurden ([Besl und R. C. Jain, 1986], [Barth et al., 1993] , [Flynn und A. K. Jain, 1989]). Beide werden im Folgenden detaillierter beschrieben und im Experimentalteil in Abschnitt 4.6.3 verglichen.

Die *numerische Schätzung* erfolgt direkt auf den 3D-Punktwolken. Mit Hilfe von Gradientenfiltern werden die partiellen Ableitungen nach u und v berechnet (siehe auch Abb. 4.31). Um zugleich eine Glättung der Datenbasis zu erreichen, wird eine modifizierte Form der gebräuchlichen Filter $[-1\ 0\ 1]$ und $[-1\ 0\ 1]^\mathsf{T}$ verwendet, die eine größere Anzahl der Nachbarpunkte in die Berechnungen einbezieht:

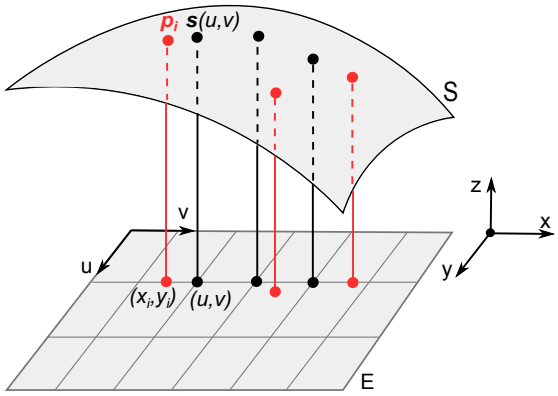
$$\mathbf{f}_v = \begin{bmatrix} -1 & -1 & -1 & 0 & 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{f}_u = \mathbf{f}_v^\mathsf{T}. \tag{4.10}$$

Aus den Vektoren $\mathbf{s}_u$ und $\mathbf{s}_v$ der partiellen Ableitungen erster Ordnung wird die erste Fundamentalform $\mathbf{I}$ gebildet:

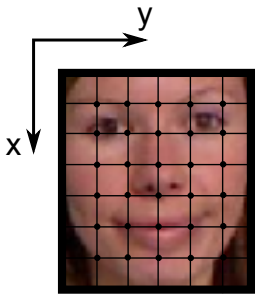
$$\mathbf{I} = \begin{bmatrix} \mathbf{s}_u \cdot \mathbf{s}_u & \mathbf{s}_u \cdot \mathbf{s}_v \\ \mathbf{s}_u \cdot \mathbf{s}_v & \mathbf{s}_v \cdot \mathbf{s}_v \end{bmatrix}. \tag{4.11}$$



(a)



(b)



(c)

Abbildung 4.30.: **(a)** Einzelne Schritte der Krümmungsschätzung und Extraktion. **(b)** Resampling der Punkte $\mathbf{p}_i$ mit Hilfe einer Referenzebene E. Für jeden xy-Gitterpunkt wird anschließend mittels linearer Interpolation ein neuer z-Wert ermittelt. **(c)** Vereinfachte Darstellung des Resamplings der xy-Ebene. Der tatsächlich gewählte Samplingabstand beträgt 2mm. Die Farbdarstellung des Gesichts dient nur der Visualisierung.

Die Ableitungen zweiter Ordnung ergeben die Elemente der zweiten Fundamentalform $\mathbf{J}$:

$$\mathbf{J} = \begin{bmatrix} \mathbf{s}_{uu} \cdot \mathbf{n} & \mathbf{s}_{uv} \cdot \mathbf{n} \\ \mathbf{s}_{uv} \cdot \mathbf{n} & \mathbf{s}_{vv} \cdot \mathbf{n} \end{bmatrix}, \tag{4.12}$$

wobei es sich bei $\mathbf{n}$ um den Normalenvektor des Oberflächenpatches handelt, der mit Hilfe des Kreuzproduktes von $\mathbf{s}_u$ und $\mathbf{s}_v$ berechnet wird [Besl und R. C. Jain, 1986]:

$$\mathbf{n}(u,v) = \frac{\mathbf{s}_u \times \mathbf{s}_v}{|\mathbf{s}_u \times \mathbf{s}_v|}. \tag{4.13}$$

Die Fundamentalmatrizen $\mathbf{I}$ und $\mathbf{J}$ kennzeichnen jedes Oberflächenpatch eindeutig und werden zur sogenannten Shape-Operator-Matrix $\mathbf{W}$ vereinigt:

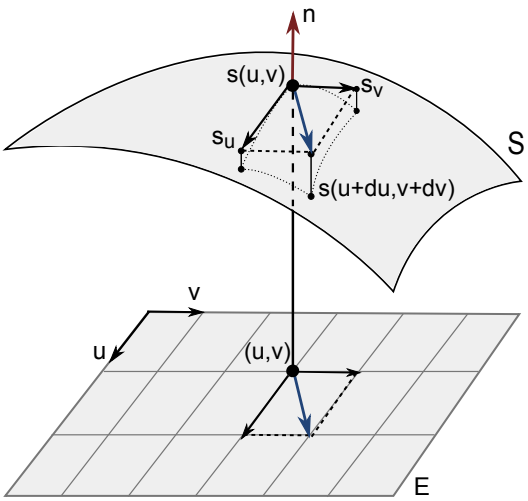


Abbildung 4.31.: Krümmungsschätzung über die Berechnung der partiellen Ableitungen nach u und v.

$$\mathbf{W} = \mathbf{I}^{-1} \cdot \mathbf{J}. \tag{4.14}$$

Aus der Shape-Operator-Matrix lassen sich die mittlere Krümmung H und die Gauß'sche Krümmung K berechnen.

$$H = \frac{1}{2} \cdot \text{spur}(\mathbf{W}), \quad K = \det(\mathbf{W}). \tag{4.15}$$

Anders als bei der numerischen Vorgehensweise erfolgt die Berechnung der Ableitungen bei der *analytischen Krümmungsschätzung* nicht direkt auf den diskreten Daten. Stattdessen werden die Nachbarpunkte durch das Fitten eines Oberflächenpatches zweiter Ordnung angenähert. Die Regression wird auf Basis der 7×7-Nachbarschaft durchgeführt, die zuvor mit ihrem Zentrumspunkt im Ursprung des Koordinatensystems registriert wurde. Da die Oberfläche einem Monge Patch entspricht, können auf Basis der Koeffizienten des quadratischen Oberflächenpolynoms:

$$z(x, y) = a_{20}\, x^2 + a_{02}\, y^2 + a_{11}\, xy + a_{10}\, x + a_{01}\, y + a_{00} \tag{4.16}$$

die Schätzungen der partiellen Ableitungen erster und zweiter Ordnung ermittelt werden [Besl und R. C. Jain, 1986]:

$$s_u = a_{10} \qquad s_v = a_{01} \qquad s_{uv} = a_{11} \qquad s_{uu} = 2 \cdot a_{20} \qquad s_{vv} = 2 \cdot a_{02}. \tag{4.17}$$

Die Gauß'sche Krümmung K und die mittlere Krümmung H ergeben sich dann aus

den folgenden Gleichungen[11]:

$$K = \frac{s_{uu}s_{vv} - s_{uv}^2}{(1 + s_u^2 + s_v^2)^2} \tag{4.18}$$

$$H = \frac{s_{uu} + s_{vv} + s_{uu}s_v^2 + s_{vv}s_u^2 - 2s_u s_v s_{uv}}{2(1 + s_u^2 + s_v^2)^{3/2}} \tag{4.19}$$

Nachdem die Werte von H und K auf numerische oder analytische Weise berechnet wurden, lassen sich mit Hilfe der HK-Klassifikation die zugehörigen Krümmungskategorien bestimmen. Ein entsprechendes Beispielergebnis wurde bereits in Abbildung 4.29b gezeigt.

Aufgrund numerischer Instabilitäten bei realen, verrauschten Daten traten die Belegungen $H = 0$ und $K = 0$ nicht auf. Daher ergaben sich bei der HK-Klassifikation nur vier der acht möglichen Kategorien aus der Tabelle 4.4. Zur Lösung wurde für alle n' Pixel eine grenzwertbasierte Filterung der geschätzten Krümmungen ergänzt:

$$H_i = \begin{cases} H_i, & |H_i| \geq \epsilon_H \\ 0, & sonst \end{cases} \tag{4.20}$$

und

$$K_i = \begin{cases} K_i, & |K_i| \geq \epsilon_K \\ 0, & sonst, \end{cases} \tag{4.21}$$

wobei gilt $i = 1...n'$. Konkrete Werte für ϵ_H und ϵ_K waren in der recherchierten Literatur nicht aufgeführt, weshalb auf Basis einer selbstaufgenommenen, planaren Ebene eigene Grenzwerte ermittelt wurden. Tiefergehende Details finden sich im folgenden Experimentalabschnitt.

Abschließend werden die Krümmungswerte, entsprechend der Ausführungen im Abschnitt 4.6.1, in Histogramme unterteilt und zu einem Merkmalsvektor konkateniert.

4.6.3. Experimentelle Untersuchung

Bei der Krümmungsanalyse handelt es sich um ein umfangreiches Themengebiet mit vielfältigen Variationsmöglichkeiten, unter anderem hinsichtlich Vor- und Nachver-

[11] Es ist zu beachten, dass die Ergebnisse der partiellen Ableitungen (s_u, s_v, s_{uu}, s_{vv}, s_{uv}) in den Gln. 4.17 bis 4.19 Skalare sind. In den Gln. 4.11 bis 4.13 treten sie als Vektoren auf. Die gekürzte Variante der analytischen Berechnungsform basiert auf der Oberflächenrepräsentation als Monge Patch und ist prinzipiell auch für die numerische möglich. Aus Gründen der Übersichtlichkeit wurde an dieser Stelle darauf verzichtet. Details dazu finden sich in [Besl und R. C. Jain, 1986].

arbeitungsschritten, Deskriptoren und Schätzverfahren. Im Folgenden werden drei Aspekte experimentell untersucht. Diese umfassen die Unterteilung des Gesichts in mehrere Extraktionsregionen, sowie die Erweiterung um eine grenzwertbasierte Filterung von H und K. Abschließend werden jeweils ein numerisches und ein analytisches Schätzverfahren gegenübergestellt.

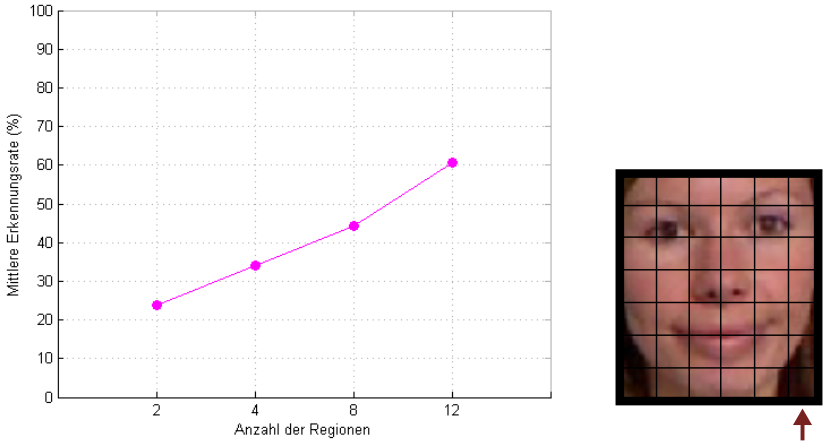
Unterteilung des Gesichts in Extraktionsregionen

Zur Festlegung der Extraktionsregionen wird das Gesicht nach einem bestimmten Schema in r Regionen unterteilt. Bei den recherchierten, patchbasierten Verfahren dominiert die Aufteilung anhand von gleichabständigen Gittern (siehe u.a. Abb. 4.4d und Tabelle B.3). Im Rahmen dieser Arbeit wurde zusätzlich, basierend auf den bisherigen Evaluationsergebnissen, eine ortsbezogene Anpassung der Auflösung getestet. Die Mund- und Wangenregionen werden dabei detaillierter aufgelöst, da für die Merkmalsdeskriptoren aus diesen Arealen eine höhere Mutual-Information zur Zielvariable geschätzt wurde als für Deskriptoren der Augenpartien. Nähere Informationen dazu finden sich in den Abbildungen 4.8, 4.18c und 4.26.

Insgesamt wurden vier verschiedene Varianten mit $r \in \{2, 4, 8, 12\}$ Regionen untersucht. Sie sind in der Abbildung 4.29d gezeigt. Die Schätzung der Krümmungen erfolgte numerisch und ohne grenzwertbasierte Filterung der Krümmungsdeskriptoren. Die allgemeine Vorgehensweise der Evaluation entspricht dem Testszenario aus Abschnitt 4.2. Die auf dem Zwölf-Klassen-Datensatz erzielten mittleren Erkennungsraten liegen zwischen 20,89 % und 60,79 %, wobei die Erkennungsrate mit der Anzahl der Regionen positiv korreliert (siehe Abb. 4.32a). Die positive Korrelation setzt sich auch bei einer weiteren Erhöhung der Regionenanzahl fort (siehe letzter Abschnitt dieses Unterkapitels 4.6.3). Je kleinteiliger die Gitteraufteilung gewählt wird, desto größer ist die Gefahr, dass einzelne Regionen ausschließlich Areale enthalten, die für die Merkmalsextraktion nicht relevant sind, wie beispielsweise Hintergründe oder Haare. Die Abbildung 4.32b verdeutlicht diesen Fall. Bei wachsender Regionenanzahl ist daher eine ergänzende Lokalisierung der elliptischen Gesichtskonturen sinnvoll, um leere Regionen zu erkennen und auszuschließen.

Festlegung eines Grenzwertes

Beim Einsatz von realen Daten ist vor der Durchführung der HK-Klassifikation eine grenzwertbasierte Filterung der Krümmungen sinnvoll. Andernfalls bleiben aufgrund des Tiefenrauschens die Bedingungen $H = 0$ und $K = 0$ unerfüllt und bestimmte Krümmungsklassen entfallen, wie die Abbildungen 4.33a und 4.33d veranschaulichen.



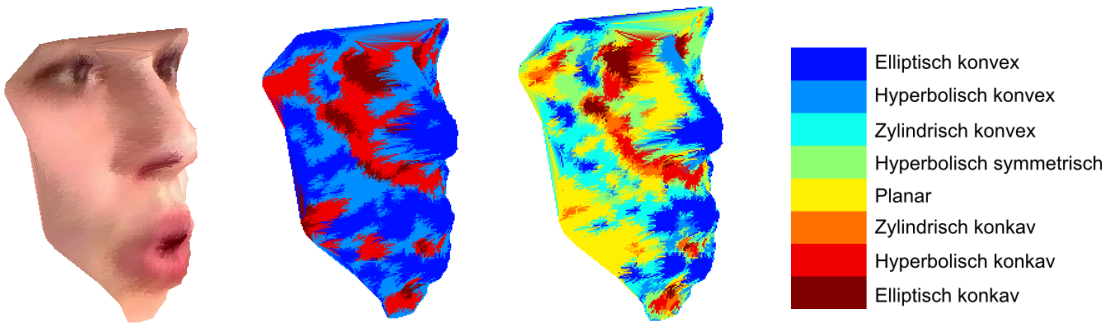
(a) (b)

Abbildung 4.32.: **(a)** Ergebnisse für verschiedene Unterteilungen des Gesichts mit $k \in \{2, 4, 8, 12\}$. **(b)** Die Unterteilung des Gesichts anhand eines 6×7 Gitters führt zu einer leeren Region (rechts unten).

Da sich in der recherchierten Literatur keine konkreten Angaben für die Grenzwerte ϵ_H und ϵ_K fanden, war die Bestimmung eigener Werte erforderlich. Dazu wurde eine planare Wand aus einer mittleren Distanz von 1,03 m aufgenommen und entsprechend der Vorverarbeitungsschritte aus Abbildung 4.30a registriert und neu abgetastet. Anschließend erfolgte die numerische Schätzung der Krümmungen H_i und K_i, mit $i = 1...n'$. Da aus theoretischer Sicht alle n' Krümmungswerte einer planaren Ebene den Wert 0 annehmen sollten, erscheint im ersten Moment die Festlegung der betragsmäßig maximalen Krümmungswerte $max(|H_i|)$ und $max(|K_i|)$ als Grenzwerte am sinnvollsten. Dies würde jedoch auch Ausreißer in die Berechnungen einbeziehen. Stattdessen wurden die Medianwerte m_{H_i} und m_{K_i} der Krümmungen über alle n' Punkte als Referenzwerte gewählt. Basierend auf diesen Referenzwerten, wurden drei verschiedene Belegungen abgeleitet und untersucht:

- Belegung 1: $\epsilon_H = m_{H_i} \wedge \epsilon_K = m_{K_i}$
- Belegung 2: $\epsilon_H = \frac{m_{H_i}}{2} \wedge \epsilon_K = \frac{m_{K_i}}{2}$
- Belegung 3: $\epsilon_H = \frac{m_{H_i}}{4} \wedge \epsilon_K = \frac{m_{K_i}}{4}$

Wie in der Abbildung 4.34 ersichtlich wird, führt die Einbindung von Grenzwerten zu einer Verbesserung der ursprünglichen mittlere Erkennungsrate (60,79 %). Die beste MER von 71,81 % wird für die zweite Belegung erzielt. Das die Belegungen mit $\epsilon_H < m_{H_i}$ und $\epsilon_K < m_{K_i}$ bessere Ergebnisse erzielen als die medianidentischen Grenzwerte der Belegung 1 erscheint im ersten Moment wenig nachvollziehbar, erklärt sich



(a) (b) (c) (d)

Abbildung 4.33.: **(a)** Punktwolke in RGB-Farben. **(b)** Punktwolke in Falschfarbendarstellung. Die Einfärbung entspricht den ohne Grenzwert numerisch geschätzten Krümmungsklassen. **(c)** Numerisch geschätzte Krümmungsklassen nach Integration einer grenzwertbasierten Filterung von H und K. **(d)** Farblegende der Krümmungsklassen.

jedoch durch die voneinander abweichenden Aufnahmeabstände der Testobjekte (siehe Kap. 3). Die Tiefenauflösung verringert sich mit wachsendem Aufnahmeabstand. Sie ist daher für die Ebene, welche aus einer mittleren Distanz von 1,03 m aufgenommen wurde, geringer als für die Gesichter des Datensatzes, deren Abstände zur Kamera zwischen 0,6 und 0,9 Metern liegen.

Numerisches vs. analytisches Schätzverfahren

Die bisherigen Ergebnisse stützen sich auf numerisch geschätzte Krümmungswerte. Zum Vergleich wurden die drei besten Konfigurationen aus der Abbildung 4.34 unter Verwendung der analytischen Krümmungsschätzung wiederholt. In allen drei Fällen erzielte die analytische Krümmungsschätzung schlechtere Erkennungsraten als die numerische, wie Abbildung 4.35 zeigt.

Detaillierte Evaluation der besten Testkonfigurationen

Die folgende, detaillierte Evaluation basiert auf der Testkonfigurationen, die in den vorhergehenden Abschnitten die beste mittlere Erkennungsrate erzielte. Sie umfasst zwölf Extraktionsregionen, ein numerisches Schätzverfahren und Grenzwerte entsprechend *Belegung 2* (siehe Seite 97). Die resultierende mittlere Erkennungsrate liegt bei 71,81 %. In der dazugehörigen Konfusionsmatrix in Abbildung 4.36 ist die unzureichende Trennung der Klassen *Kuss*, *OForm* und *UForm* erkennbar, die sich bereits bei den anderen Merkmalstypen zeigte. Die übungsspezifischen Erkennungsraten der drei Klassen liegen zwischen 32,32 % und 54,6 %. Gleichzeitig werden jedoch

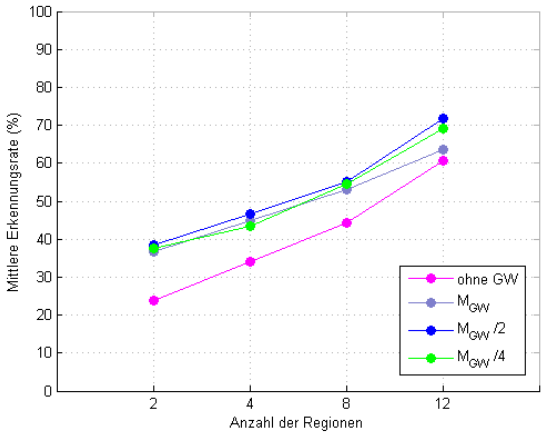


Abbildung 4.34.: Einfluss der grenzwertbasierten Filterung von H und K auf die mittlere Erkennungsrate. Beim Einsatz von 12 Extraktionsregionen verbessert sich die ursprüngliche MER von 60,79 % auf 63,50 %, 69,23 % oder 71,81 %, je nach Belegung der Grenzwerte. Die Krümmungen wurden numerisch geschätzt.

39,14 % der *UForm*-Observationen als *Kuss*-Übung klassifiziert. Die Erkennung der übrigen neun Klassen ist mit übungsspezifischen Erkennungsraten zwischen 72,13 % und 87,4 % deutlich zuverlässiger.

Bei zwölf Extraktionsregionen und Histogrammen mit acht Bins ergibt sich ein Merkmalsvektor mit $12 \times 8 = 96$ Variablen. Mit Hilfe der Mutual-Information lässt sich für jede dieser Variablen der Zusammenhang zur ausgeführten Übung schätzen. Um diese Ergebnisse übersichtlich in einem Balkendiagramm zusammenfassen zu können, wurden die acht MI-Werte einer Region zu einem einzelnen Mittelwert zusammengefasst. Das Balkendiagramm und die Positionierung der Regionen innerhalb des Gesichts sind in den Abbildungen 4.37a und 4.37b visualisiert. In Übereinstimmung zu den Ergebnissen der anderen Merkmalsevaluationen ist für die Merkmale der Mund- und Wangenregionen eine höhere MI und somit ein größerer Zusammenhang zur ausgeführten Übung sichtbar als für die Merkmale der Augenpartien.

Für die vorliegende Parameterkonstellation wurde auch eine Unterteilung des Gesichts in 14 Regionen, entsprechend der Abbildung 4.37c, untersucht. Diese führte zu einer weiteren Verbesserung der MER von 71,81 % auf 73,22 %. Weiterführende Experimente deuten darauf hin, dass durch eine noch feinere Unterteilung eine weitere Steigerung der MER möglich ist. Wie bereits beschrieben wurde ist hierfür jedoch eine weiterentwickeltere Vorgehensweise bei der Regionenaufteilung vonnöten.

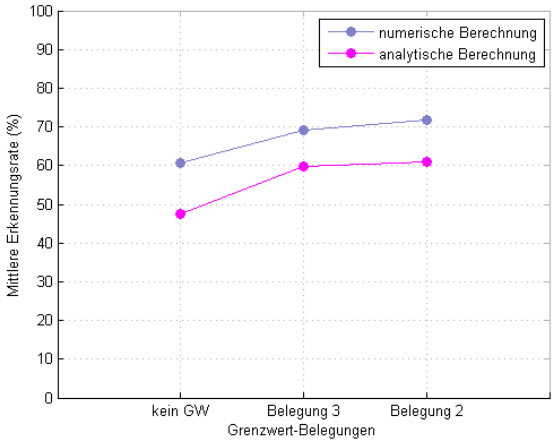


Abbildung 4.35.: Gegenüberstellung der Ergebnisse der numerischen und der analytischen Krümmungsschätzung (12 Extraktionsregionen). Eine Erläuterung der einzelnen Testbelegungen findet sich auf der Seite 103.

4.6.4. Fazit

Im Rahmen der Krümmungsmerkmalsevaluation wurden insgesamt drei Aspekte untersucht. Sie umfassen die Anzahl der Regionen, die Einbindung einer grenzwertbasierten Filterung und die Art der Krümmungsschätzung. Für die Regionenanzahl zeigte sich eine positive Korrelation mit der mittleren Erkennungsrate. Für zwei bis zwölf Regionen wurden Erkennungsraten zwischen 20,89 % und 60,79 % erzielt. Über eine zusätzliche grenzwertbasierte Filterung der Krümmungen H und K ließ sich eine weitere Verbesserung auf 71,81 % erreichen. Das analytische Schätzverfahren schnitt im Vergleich zum numerischen in dieser Arbeit schlechter ab. Die Schätzung der Mutual-Information ergab für die Merkmalsvariablen der Mund- und Wangenregion einen relativ betrachtet stärkeren Zusammenhang zur ausgeführten Übung als für die Merkmalsvariablen der Augenpartie. Dies stimmt mit den für die übrigen Merkmalsextraktionsverfahren erzielten Ergebnissen überein.

4.7. Zusammenfassung und Fazit

Die vorliegende Zusammenfassung umfasst zwei Abschnitte. Im ersten Abschnitt werden die Ergebnisse der einzelnen Merkmalsevaluationen gegenübergestellt, um abschließend einen Leitfaden für die Merkmalsauswahl zur tiefendatenbasierten Gesichtsanalyse zu erstellen.

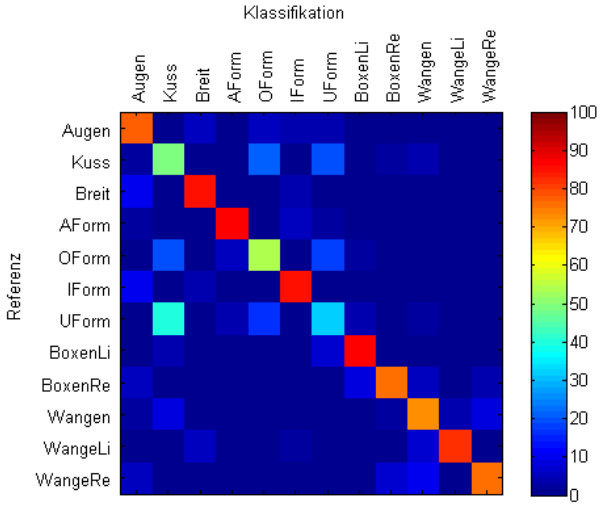


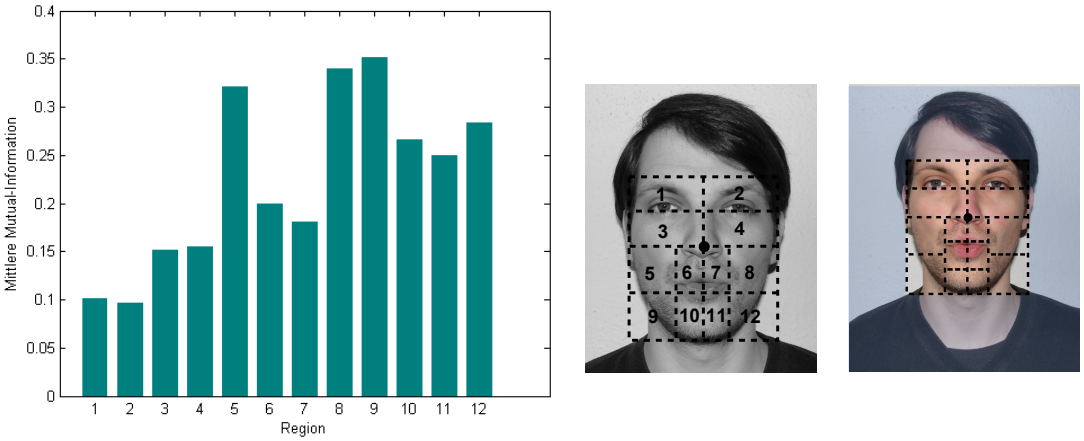
Abbildung 4.36.: Konfusionsmatrix für die Klassifikation mit Krümmungsmerkmalen, basierend auf 12 Regionen, einem numerischen Schätzverfahren und Grenzwerten entsprechend *Belegung 2*. Die MER beträgt 71,81 %.

4.7.1. Überblick über die Evaluationsergebnisse

Auf Grundlage des in Abschnitt 4.2 beschriebenen Testszenarios wurden in diesem Kapitel fünf verschiedene Merkmalsextraktionsverfahren evaluiert. Neben den *Distanz-* und *Winkelmerkmalen* umfassen diese *Punktsignaturen*, *Histogramme orientierter Normalenvektoren* und *Krümmungsmerkmale*. Die erzielten mittleren Erkennungsraten (MER) liegen bei zwölf Übungsklassen zwischen 51,08 % und 75,40 %. Das schlechteste Klassifikationsergebnis lieferten die Distanzmerkmale, das Beste die Kombination mehrerer Punktsignaturen.

Einige Beispielergebnisse aus der recherchierten Literatur finden sich in Tabelle 4.5, wobei eine pauschale Übertragbarkeit der Resultate auf die Ergebnisse dieser Arbeit nur eingeschränkt möglich ist. Da unterschiedliche Datensätze verwendet wurden, weicht die Qualität der Tiefendaten, die Art der Kategorien und die Anzahl der Klassen voneinander ab. So beinhaltet der BU-3DFE-Datensatz, der in vielen Referenzpublikationen Verwendung findet, nur sechs bzw. sieben Klassen, während in dieser Arbeit zwölf Übungsklassen unterschieden werden [Yin et al., 2006]. Desweiteren konzentriert sich der Bewegungsschwerpunkt der therapeutischen Übungen auf die Mundregion, während die gemimten Emotionen sowohl in der Mund- als auch der Augenregion charakteristische Bewegungsmuster aufweisen, wie ein Vergleich der Abbildungen 4.38 und 4.39 verdeutlicht.

Die in [Rabiu et al., 2012] beschriebene *Distanz- und Winkelmerkmalsextraktion* wurde exakt nachimplementiert und ergab für die zwölf Übungsklassen eine MER von



(a) (b) (c)

Abbildung 4.37.: **(a)** Mittlere MI der einzelnen Regionen. **(b)** Positionierung der Regionen innerhalb des Gesichts. **(c)** Eine weitere Unterteilung des Gesichts in 14 Regionen resultiert in einer Verbesserung der MER auf 73,22 %.

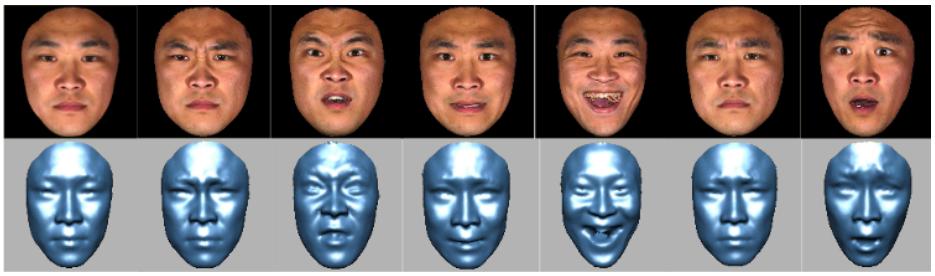
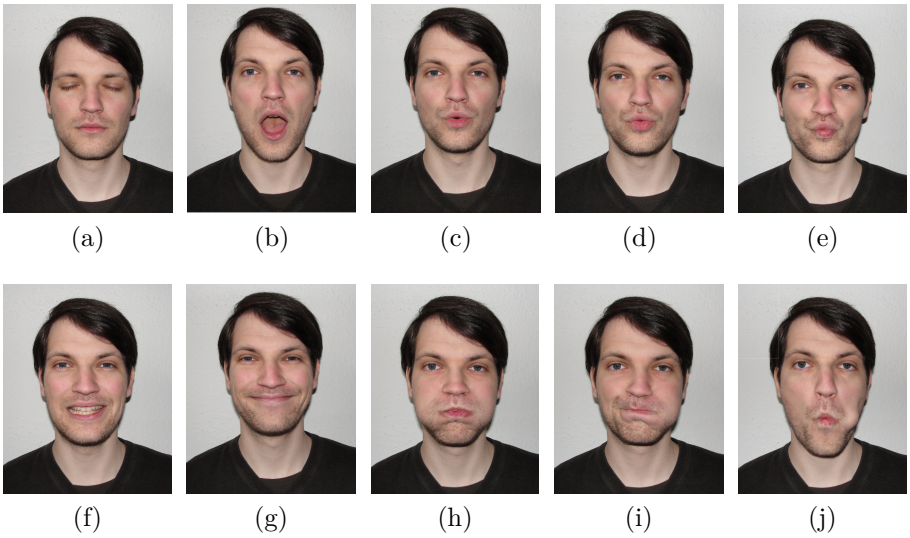


Abbildung 4.38.: Die sieben Klassen des BU-3DFE-Datensatzes bestehend aus dem Neutralgesicht, sowie den sechs Basisemotionen (Wut, Ekel, Angst, Freude, Trauer, Überraschung). Bildquelle: [Yin et al., 2006].

60,62 %. Die Experimente der Referenzveröffentlichung basieren auf dem BU-3DFE-Datensatz und erzielten für die Klassifikation in sieben Basisemotionen eine MER von 92,2 %. Die Ergebnisse gelten für konkatenierte Merkmalsvektoren, bestehend aus der Distanz- und Winkelinformation. Eine getrennte Evaluation der extrahierten Distanzen und Winkel findet sich in [Rabiu et al., 2012] nicht, wurde in dieser Arbeit jedoch ergänzend durchgeführt. Die Winkelmerkmale erzielten mit 64,16 % eine deutlich höhere MER als die Distanzmerkmale allein (51,08 %) und schnitten zugleich besser ab als die konkatenierte Variante. Anders als die Winkel werden die aus der linken und rechten Gesichtshälfte extrahierten Distanzen teilweise gemittelt und auf diese Weise zu einer Merkmalsvariable zusammengefasst. Das Weglassen dieses Schrittes wäre ein sinnvoller Anknüpfungspunkt für den Versuch, eine weitere Verbesserung der Erkennungsrate zu erreichen.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Abbildung 4.39.: Die in dieser Arbeit unterschiedenen zwölf therapeutischen Übungsklassen. **(a)-(e)** *Augen*, *AForm*, *OForm*, *UForm*, *Kuss*. **(f)-(j)** *IForm*, *Breit*, *Wangen*, *WangeLi*, *BoxenLi*. Zu den Übungen *WangeLi* und *BoxenLi* existiert jeweils eine achsengespiegelte Variante.

Tabelle 4.5.: Beispielergebnisse aus der recherchierten Literatur zur tiefendatenbasierten Gesichtsanalyse (siehe auch Tab. B.3 im Anhang).

Quelle	**MER**	**Klassen**	**Datensatz**	**Erläuterungen**
[Rabiu et al., 2012]	92,2 %	7	BU-3DFE	Konkatenierte Distanz- und Winkelmerkmale. Das Verfahren wurde in dieser Arbeit nachimplementiert.
[J. Wang et al., 2006]	max. 83,6 %	6	BU-3DFE	Histogramme über 12 Krümmungskategorien.
[Lemaire et al., 2013]	76,6 %	6	BU-3DFE	Krümmungsmerkmale, kodiert mit HOGs.
[Y. Wang et al., 2013]	76,56 %	6	Bosphorus	Verschiedene Krümmungsmerkmale, jeweils codiert mit LBP.

Das Ziel der *Punktsignaturextraktion* in [Chua et al., 2000] besteht in der Personenverifizierung. Diese basiert auf einer sogenannten Voting-Rate, die sich nicht direkt mit der MER vergleichen lässt. Zudem wurde das Originalverfahren geringfügig modifiziert, um es auf die Übungsklassifikation anwenden zu können. Insgesamt beinhaltet die Analyse in dieser Arbeit die Evaluation von 24 einzelnen Punktsignaturen, sowie von drei konkatentierten Varianten. Jede einzelne Punktsignatur ist durch einen von drei Abtastwinkeln $\Delta\theta \in \{5{,}7°; 11{,}6°; 24°\}$, sowie einen von acht Radien mit $r \in \{4\text{cm}; 4{,}5\text{cm}; 5\text{cm}; 5{,}5\text{cm}; 6\text{cm}; 6{,}5\text{cm}; 7\text{cm}; 7{,}5\text{cm}\}$, definiert. Die konkatenierten Merkmalsvektoren bestehen aus den aneinandergefügten Punktsignaturen aller acht Radien r für ein festgelegtes Abtastintervall $\Delta\theta$. Wie in der Abbildung 4.40 ersichtlich ist, sind die beiden besten mittleren Erkennungsraten der einzelnen Punktsignaturen (65,74 %, 64,57 %) vergleichbar mit der MER der Winkelmerkmale (64,16 %). Mit 64 bzw. 32 statt 27 Dimensionen umfassen sie jedoch auch einen größeren Merkmalsraum. Die Konkatenierung von mehreren Punktsignaturen erzielte mit 75,40 % und 73,89 % einen deutlichen Anstieg der Erkennungsraten, resultierte jedoch auch in Merkmalsraumdimensionen von 512 bzw. 256. Der Vorteil der Punktsignaturen gegenüber den Distanz- und Winkelmerkmalen besteht zusätzlich darin, dass die Anforderungen an eine vorgeschaltete Landmarkenlokalisierung deutlich geringer sind, da nur zwei Landmarken, nämlich ein Zentrum und ein Referenzpunkt der Signatur, bestimmt werden müssen.

Die Extraktion und Klassifikation von *Histogrammen orientierter Normalenvektoren* erzielte mit einer MER von 63,41 % ebenfalls vergleichbare Ergebnisse wie der Einsatz der Winkelmerkmale, führte mit $5 \times 5 \times 8 = 200$ Dimensionen jedoch zu einem deutlich größeren Merkmalsraum. Mit Hilfe der linearen Diskriminanzanalyse ließ sich die Zahl der Dimensionen auf $k - 1$ reduzieren, wobei $k = 12$ der Anzahl der Klassen entspricht. Die mittlere Erkennungsrate verbesserte sich dadurch auf 66,79 %.

Die *Krümmungsmerkmale* erzielten mit einer vergleichsweise geringen Anzahl an Merkmalsvariablen ähnliche Ergebnisse wie die konkatenierten Punktsignaturen. So ergibt sich bei zwölf Extraktionsregionen ein Merkmalsvektor von $8 \times 12 = 96$ Einträgen und eine MER von 71,81 %. Eine Erhöhung auf vierzehn Regionen führt zu einem 112-dimensionalen Merkmalsvektor und einer MER von 73,22 %. Das implementierte Extraktionsverfahren vereinigt Aspekte aus verschiedenen Publikationen (u.a. [Besl und R. C. Jain, 1986], [J. Wang et al., 2006], [Colombo et al., 2006]). Unter anderem aus diesem Grund dienen die Beispielraten der anderen krümmungsbasierten Verfahren, welche zwischen 76,56 % und 83,6 % liegen, nur der ungefähren Orientierung (siehe Tabelle 4.5). Zudem ist auch an dieser Stelle zu beachten, dass die Ergebnisse auf unterschiedlichen Datensätzen beruhen.

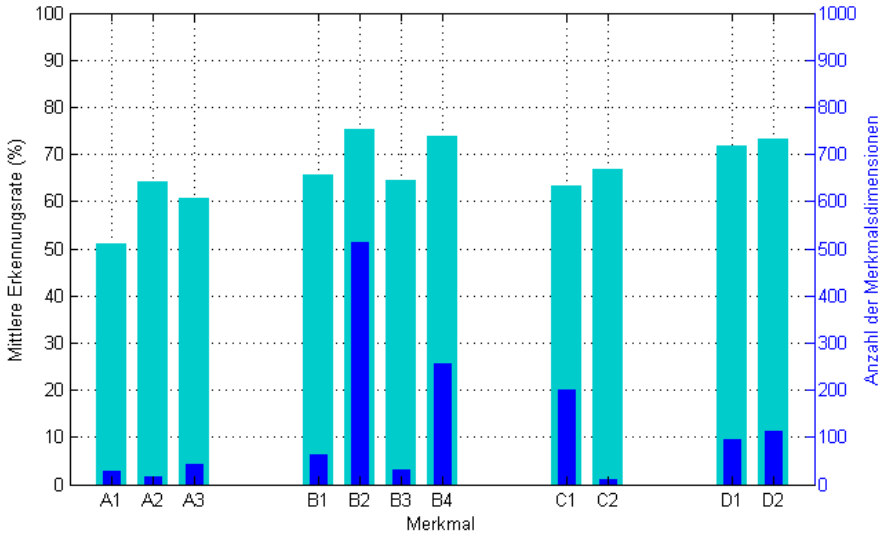


Abbildung 4.40.: Gegenüberstellung der einzelnen Merkmalstypen hinsichtlich der mittleren Erkennungsrate (hellblau) bei zwölf Klassen und der Anzahl der Merkmalsdimensionen (dunkelblau). Das Säulendiagramm enthält eine Auswahl der besten Ergebnisse für jeden Merkmalstyp. (*A1*: Distanzmerkmale, *A2*: Winkelmerkmale, *A3*: Konkatenierung der Distanz- und Winkelmerkmale, *B1*: Einzelne Punktsignatur mit $\Delta\theta = 5{,}7°$ und $r = 6{,}5$cm , *B2*: Konkatenierte Punktsignaturmerkmale über alle acht r mit $\Delta\theta = 5{,}7°$, *B3*: Einzelne Punktsignatur mit $\Delta\theta = 11{,}6°$ und $r = 6$cm, *B4*: Konkatenierte Punktsignaturmerkmale über alle acht r mit $\Delta\theta = 11{,}6°$, *C1*: Histogramme orientierter Normalenvektoren (12 Regionen, 5×5 Bins umfassendes, bivariates Histogramm), *C2*: Histogramme orientierter Normalenvektoren wie in *C1*, ergänzt durch eine Merkmalsraumreduktion mittels LDA, *D1*: Krümmungsanalyse (12 Regionen) , *D2*: Krümmungsanalyse (14 Regionen)).

4.7.2. Leitfaden für die Merkmalsauswahl

Neben den Klassifikationsergebnissen und der Dimensionsanzahl, welche Gegenstand des vorhergehenden Abschnitts waren, sind auch andere Aspekte für die Merkmalswahl entscheidend. So kann die Notwendigkeit einer vorgeschalteten Landmarkenlokalisierung eine zusätzliche Fehlerquelle darstellen und somit den Einsatz in einem realen Szenario erschweren. Diese Problematik betrifft insbesondere die Distanz- und Winkelmerkmale, deren Extraktion auf 29 Landmarken basiert. Für die Punktsignaturmerkmale sind hingegen nur zwei Landmarken erforderlich. Bei diesen handelt es sich zum einen um einen Referenzpunkt, welcher den Anfang einer Signatur kennzeichnet, sowie zum anderen um die Nasenspitze, welche im Zentrum der Signatur liegt und sich aufgrund ihrer charakteristischen Krümmung robust erkennen lässt. Für die Ex-

traktion der HON- und Krümmungsmerkmale genügt im Prinzip eine Bounding-Box. Sie ist das Ergebnis einer vorgeschalteten Gesichtsdetektion und wird mit Hilfe eines Gitters in mehrere Extraktionsregionen unterteilt. Dieses Gitter wird in den ausgewerteten Publikationen nach einem festen Schema in $m \times n$ Felder unterteilt und ist somit unabhängig von einzelnen Landmarken (siehe Tabelle B.3). In dieser Arbeit dient die Nasenspitze als zusätzlicher Ankerpunkt für die Gitteraufteilung, um eine weitgehend konsistente Zuordnung der Gitterfelder zu den Regionen des Gesichts auch dann sicherzustellen, wenn sich die relativen Proportionen, beispielsweise durch weites Öffnen des Mundes, ändern.

Der landmarkenbasierte Charakter der Distanz- und Winkelmerkmale führt zu einem weiteren Nachteil. So werden homogene, landmarkenarme Gesichtsbereiche, wie beispielsweise die Wangenregion, bei der Merkmalsextraktion nicht erfasst. Dies führt zu deutlich schlechteren Erkennungsraten für die Wangenübungen *BoxenLi*, *BoxenRe*, *Wangen*, *WangeLi* und *WangeRe* (vergleiche dazu die Konfusionsmatrizen in den Abbildungen 4.10, 4.19, 4.25b und 4.36). Für weiterführende Experimente wäre es zudem sinnvoll, dass, gemäß dem Ansatz von [Rabiu et al., 2012] durchgeführte, Zusammenfassen der aus der linken und rechten Gesichtshälfte extrahierten Distanzen nicht durchzuführen und diese stattdessen einzeln in den Merkmalsvektor einzubinden (z.B. $d22$ und $d23$ bei δ_{10}).

Bei der Extraktion der patchbasierten HON- und Krümmungsmerkmale empfiehlt sich eine Unterteilung des Gesichts in mehrere Extraktionsregionen, damit der Ortsbezug der extrahierten Informationen beibehalten wird. Dies ist beispielsweise durch eine Histogrammbildung für jede Region mit anschließender Konkatenierung der regionenspezifischen Merkmalsvektoren möglich. Bei den Punktsignaturen entspricht dies der Konkatenierung mehrerer Punktsignaturen mit unterschiedlichen Radien r.

Der kombinierte Einsatz der Vektoren aller fünf Merkmalstypen führt zu einer weiteren deutlichen Verbesserung der Erkennungsraten. Für die detaillierte Auswertung sei auf den Abschnitt 5.3.2 verwiesen.

5. Feedbackgenerierung und Implementierung des Prototypen

Das vorliegende Kapitel unterteilt sich in zwei thematische Schwerpunkte. Der erste Teil erstreckt sich über die Unterkapitel 5.1 bis 5.3 und befasst sich mit der Entwicklung eines Verfahrens zur Feedbackgenerierung. Dieses bildet den letzten Prozessschritt der technischen Gesamtarchitektur, welche in reduzierter Form in der Abbildung 5.1 gezeigt ist. Die vollständige Übersicht findet sich in der Abbildung 6.1. Der zweite Themenkomplex dieses Kapitels beschreibt die Entwicklung des, auf der technischen Gesamtarchitektur basierenden, Prototypen. Beide Themenschwerpunkte werden jeweils durch experimentelle Analysen vervollständigt.

5.1. Eingrenzung von geeigneten Methoden zur Feedbackerzeugung

Wie die Abbildung 5.1 verdeutlicht, basiert der Prozess der Feedbackerzeugung auf den zuvor extrahierten Merkmalsdeskriptoren. Das angestrebte Resultat der Feedbackerzeugung ist ein diskretes oder kontinuierliches Bewertungsmaß zur Beurteilung der Übungsausführung des Patienten. In den folgenden Abschnitten werden verschiedene Möglichkeiten der Datenanalyse vorgestellt und diskutiert, inwieweit diese eine Quantifizierung der Ausführungsqualität, auf Basis der extrahierten Merkmalsdeskriptoren, erlauben. Zu den klassischen Verfahren und Werkzeugen zählen dabei die Klassifikation, die Regression und die Berechnung von Abstandsmaßen (Abschn. 5.1.1, 5.1.3 und 5.1.4). Ein Exkurs zum Facial-Action-Coding-System findet sich in Abschnitt 5.1.2 und ergänzt die Diskussion zum Einsatz eines klassifikationsbasierten Ansatzes.

5.1.1. Klassifikation

Ein Klassifikator ordnet einem Anfragebild eines vorliegenden Datensatzes auf Basis der Ground-Truth und der extrahierten Merkmalsdeskriptoren $\mathbf{x}$ eine diskrete Klasse zu ([Duda et al., 2001], [Bishop, 2006]). In dieser Arbeit handelt es sich dabei um

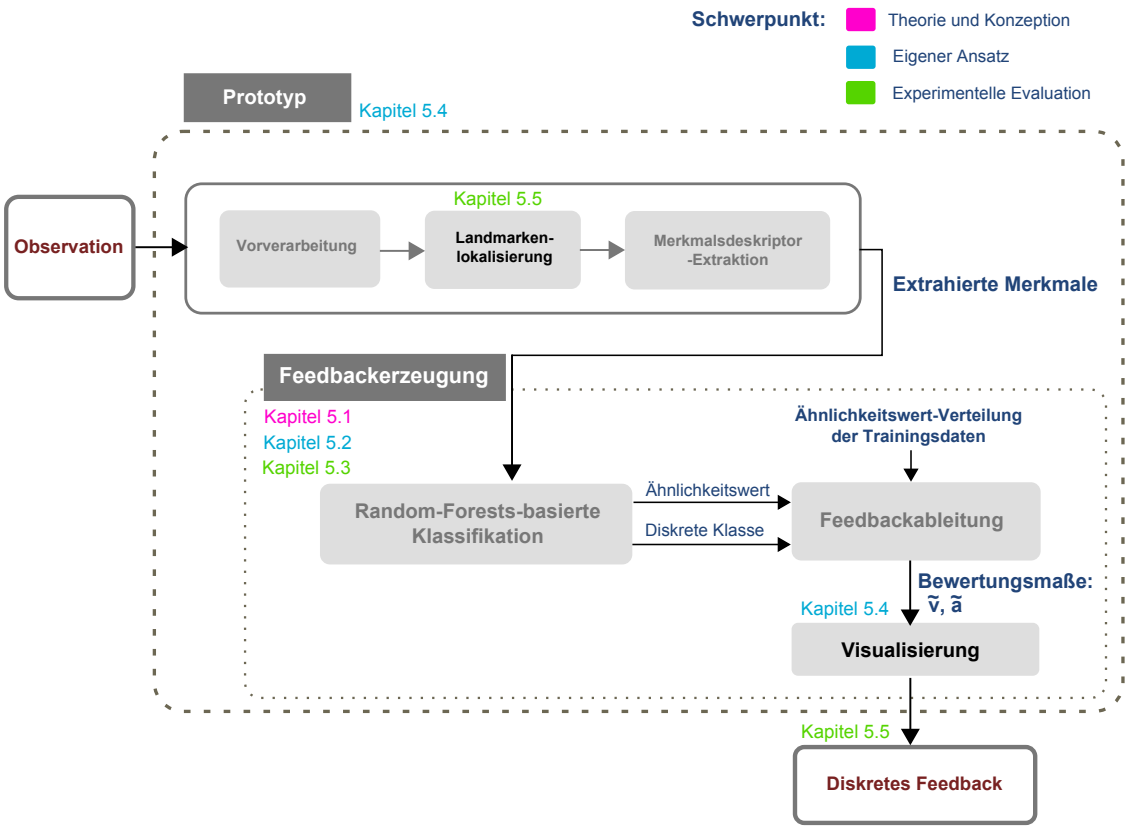


Abbildung 5.1.: Schematische Übersicht über die technische Gesamtarchitektur des implementierten Prototypen. Die Feedbackerzeugung bildet den letzten Prozessschritt und basiert auf den extrahierten Merkmalswerten.

eine von zwölf therapeutischen Fazialisübungen ω_i, mit $i \in \{1, ..., 12\}$. Detailliertere Informationen zu den Übungen finden sich in Abschnitt 2.3.1. Einige Klassifikatortypen ermöglichen zudem die Ermittlung sogenannter *A-posteriori-Wahrscheinlichkeiten* ([Duda et al., 2001], [Bishop, 2006]). Diese stellen Schätzwerte für die Auftrittswahrscheinlichkeit einer Klasse ω_i bei vorliegendem Merkmalssatz $\mathbf{x}$ dar. Die möglichen Werte, die $P(\omega_i|\mathbf{x})$ annehmen kann, sind kontinuierlich und liegen im Intervall $[0\,\%, 100\,\%]$. Eine Wahrscheinlichkeit im oberen Bereich des Intervalls deutet auf eine hohe Übereinstimmung mit dem trainierten klassenspezifischen Modell hin und *kann* ein Hinweis auf eine korrekte Übungsausführung sein. Als Bewertungsmaß für die Qualität der Übungsausführung ist $P(\omega_i|\mathbf{x})$ dennoch nicht geeignet. Zur Verdeutlichung dessen werden im Folgenden zwei Beispiele aufgeführt. Ausgangspunkt beider Beispiele ist ein Übungsszenario, bei welchem ein Patient A eine Zielübung ω_1 ausführen soll. Mit Hilfe eines trainierten Klassifikators und den extrahierten Merkmalsdeskriptoren $\mathbf{x}_A$ wird im Anschluss die A-posteriori-Wahrscheinlichkeit $P(\omega_1|\mathbf{x}_A)$ berechnet.

- *Beispiel 1:* Würde die Zielübung ω_1 beispielsweise der Übung *Kuss* entspre-

chen, wären die Ergebnisse $P(\omega_1|\mathbf{x}_A) = 20\,\%$ und $P(\omega_j|\mathbf{x}_A) = 80\,\%$, mit $j \neq 1$, aus Sicht eines menschlichen Betrachters unterschiedlich einzuschätzen, je nachdem, ob die Klasse ω_j die Fazialisübung *UForm* oder *Breit* repräsentiert (vgl. Abb. 5.2). Eine niedrige A-posteriori-Wahrscheinlichkeit für die Soll-Übung ω_1 deutet somit nicht zwingend auf eine gänzlich unzulängliche Übungsausführung hin, sondern ist immer im Kontext der Ähnlichkeiten zwischen den Übungsklassen zu betrachten. Wie stark die Distanz und Überschneidung bestimmter Klassencluster im Merkmalsraum ausgeprägt ist, ist nicht zuletzt von der Wahl der zu extrahierenden Merkmale abhängig. Die auf Basis eines Testdatensatzes ermittelten Konfusionsmatrizen können diesbezüglich erste Hinweise liefern (vgl. Abbildungen 4.10, 4.19, 4.25b und 4.36).

- *Beispiel 2:* Eine A-posteriori Wahrscheinlichkeit von beispielsweise $P(\omega_1|\mathbf{x}_A) = 80\,\%$, könnte zunächst vermuten lassen, dass die Übungsausführungsqualität der Testperson sehr positiv einzuschätzen ist, da sie im oberen Bereich des Intervalls $[0\,\%, 100\,\%]$ liegt. Die Abbildung 5.3 verdeutlicht jedoch beispielhaft die besondere Bedeutung der klassenspezifischen trainingsdateninternen Distanzen. Sind die Intraklassendistanzen mehrheitlich geringer als die Abstände zur Testobservation, deutet dies auf eine unzureichende Übungsausführung hin.

Wie in den beschriebenen Beispielen deutlich wurde, ist ein auf der Fazialisübungsklassifikation basierender Ansatz allein nicht für die Entwicklung eines Feedbackssystems geeignet. Das im folgenden Abschnitt beschriebene Facial-Action-Coding-System umgeht die beschriebenen Nachteile durch detailliertes Expertenwissen in der Ground-Truth.

5.1.2. Facial Action Coding System

Mit Hilfe des *Facial-Action-Coding-Systems* (FACS) lässt sich die Mimik eines Gesichts durch eine oder mehrere *Action-Units* (AU) beschreiben ([Ekman und Friesen, 1976], [Cohn et al., 2007]). Jeder Action-Unit sind spezifische Muskelbewegungen zugeordnet, wobei zusätzlich zwischen fünf verschiedenen Intensitätsstufen unterschieden wird. Da auch jede der zwölf therapeutischen Mimikübungen aus mindestens einer AU besteht, ließe sich die korrekte Übungsausführung in Form einer Checkliste überprüfen. Im Rahmen des Feedbacks würde der Patient auf nicht oder inkorrekt ausgeführte AUs hingewiesen werden. Eine inkorrekte Ausführung würde bei zu schwacher oder zu intensiver Aktivierung der AU vorliegen. Der Ansatz wurde in dieser Arbeit nicht gewählt, da die Erstellung einer anerkannten Ground-Truth nur durch zerti-

Abbildung 5.2.: Zeilenweise Sortierung von Fazialisübungen nach beispielhaften, übereinstimmenden Eigenschaften in der unteren Gesichtshälfte. **Zeile 1:** Die konkave Wangenwölbung der Übung *Wangen* findet sich bei den Übungen *WangeLi* und *WangeRe* wieder (Festlegung von rechts und links gemäß Patientensicht). **Zeile 2:** Aktivierung des Mundringmuskels bei Ausführung der Übungen *OForm*, *UForm* und *Kuss*. **Zeile 3:** Laterale Verschiebung der Mundwinkel (Übungen *IForm* und *Breit*) (Bild 1-2). Die Übung *AForm* hebt sich durch den weit geöffneten Mund stärker von den anderen Übungen ab (Bild 3).

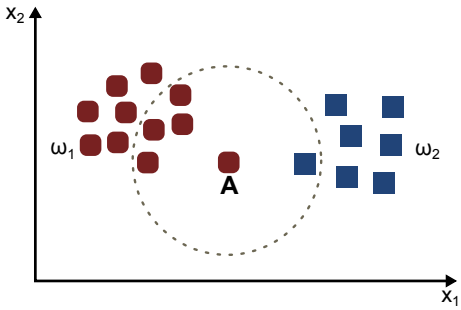


Abbildung 5.3.: Beispielhafte Visualisierung der Trainingsobservationen von zwei Klassen ω_1 und ω_2 in einem zweidimensionalen Merkmalsraum, welcher von den Merkmalsvariablen x_1 und x_2 aufgespannt wird. Die dem Patienten A zugewiesene Testobservation liegt im Zentrum des eingezeichneten Kreises. Im Fall einer kNN-Klassifizierung, mit $k = 5$, würde für die Testobservation eine A-posteriori-Wahrscheinlichkeit von $P(\omega_1|\mathbf{x}_A) = 80\,\%$ geschätzt werden. Relativ gesehen ist dieser Wert gering, da sich für die Trainingsobservationen der Klasse ω_1 im Fall eines Leave-One-Out-Testszenarios jeweils A-priori-Wahrscheinlichkeiten von $100\,\%$ ergeben würden.

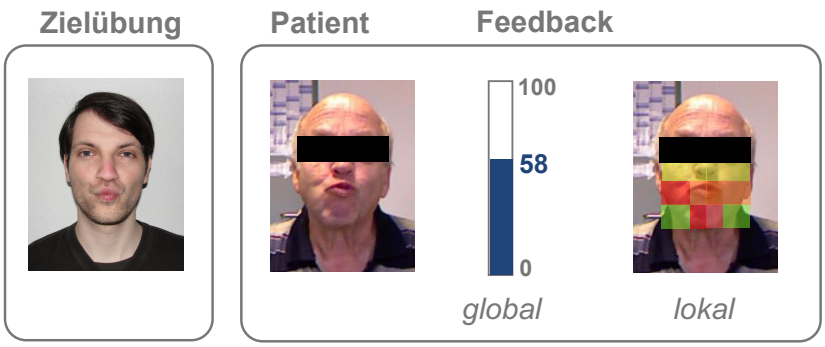


Abbildung 5.4.: Beispielhafte Visualisierung von regressionsbasiertem globalen und lokalem Feedback. Im globalen Fall, mit $k = 1$, dient eine kontinuierliche Skalenanzeige zur Darstellung. Für $k = 10$ lokale Feedbackregionen wird das Intervall $[0; 100]$ auf einen Farbverlauf von rot bis grün abgebildet. Sowohl der Skalenwert $v = 58$, als auch die farbliche Darstellung, wurden willkürlich festgelegt und dienen ausschließlich der Demonstration.

fizierte FACS-Kodierer möglich ist. Um die Reliabilität der Datenbasis zu sichern, sind zudem mindestens zwei Kodierer erforderlich [Cohn et al., 2007]. Ein verfügbarer FACS-Datensatz, der allen Anforderungen dieser Arbeit entspricht (Tiefendaten, Fazialisübungen), konnte, trotz sorgfältiger Suche, im Rahmen dieser Arbeit nicht ausfindig gemacht werden.

5.1.3. Regression

Anders als die Klassifikation, ordnet die Regression einem Anfragebild keine diskrete Klasse sondern einen kontinuierlichen Ausgabewert zu. Um eine Regressionsfunktion trainieren zu können, deren Ausgabewert ein Maß für die Qualität der Übungsdurchführung darstellt, wird eine entsprechende Ground-Truth benötigt. Jedem Trainingsbeispiel der Ground-Truth muss dabei ein Wert v, mit $v \in [0; 100]$, zugeordnet sein, wobei 0 eine ungenügende und 100 eine sehr gute Ausführung kennzeichnet. Wird für s verschiedene Teilregionen gesondertes Feedback gewünscht, müssen jeder Trainingsobservation entsprechend s Werte zugeordnet werden. Beispiele dafür, wie sich diese Ausgabewerte für globales ($s = 1$) bzw. lokales ($s = 10$) Feedback visualisieren ließen, finden sich in der Abbildung 5.4.

Die Zuordnung von v zu den einzelnen Trainingsdaten wird auch als Annotation bezeichnet und erfolgt in der Regel manuell. Sie sollte durch qualifiziertes Fachpersonal, wie Logopäden oder Sprechwissenschaftler, vorgenommen werden. Der manuelle Annotationsvorgang weist jedoch Nachteile auf. Zum einen ist er zeit- und personalin-

tensiv, insbesondere, weil für jede Übung eine große Anzahl repräsentativer Trainingsdaten benötigt werden. Zum anderen ist er den Einflüssen subjektiver Einschätzungen unterworfen. Die Elektromyografie ist als Hilfsmittel zur Objektivierung des Annotationsverfahrens nicht geeignet, da sie zu Verdeckungen innerhalb des Gesichts führt. Dies würde die nachfolgende Extraktion der Tiefenmerkmale beeinträchtigen. Auch objektive Bewertungsmaße wie Winkel oder Abstände innerhalb des Gesichts sind zur Unterstützung des Annotierenden nur eingeschränkt einsetzbar, da für die meisten Fazialisübungen eindeutige und subjektübergreifende Anfangs- und Zielzustände nur schwer definierbar sind. Deutlich wird dies bei einem Vergleich mit der in Abbildung 5.5a gezeigten physischen Mobilisierungsübung. Dabei beschreibt α den Winkel zwischen Arm- und Rumpf. Der Fortschritt des Patienten beim Heben des Armes lässt sich über das Verhältnis $\alpha/90°$ quantifizieren. Da der komplexe Aufbau der Gesichtsmuskulatur jedoch eine Vielzahl, auch voneinander unabhängiger, Mimikbewegungen erlaubt, lässt sich ein vergleichbares metrisches Referenzmaß deutlich schwerer bestimmen. So wäre beispielsweise zur Bewertung der Ausführung der *Kuss*-Übung ein definierter Zielabstand zwischen der Mundspitze und den Zähnen wenig intuitiv. Die Abbildungen 5.5b bis 5.5e verdeutlichen die Ausführungsunterschiede der *Kuss*-Übung zwischen einer gesunden Person und drei Fazialisparesepatienten. Eine weiterer Nachteil von metrischen Zieldefinitionen besteht darin, dass sich Distanzen und Winkel am verlässlichsten auf Basis von markanten Referenzpunkten ermitteln lassen. In landmarkenarmen, homogenen Arealen, wie beispielsweise den Wangenregionen, fehlen entsprechende Orientierungspunkte (vgl. auch Abschn. 4.3.2).

Zuletzt ist zu bedenken, dass ein kontinuierliches Bewertungsmaß den Eindruck eines sehr fein abgestuften Feedbacks erzeugt. Es bliebe kritisch zu hinterfragen, ob manuelle Ground-Truth-Annotationen diesen Anspruch erfüllen können. So umfassen Action-Units, welche von unterwiesenen FACS-Kodierern gesetzt werden, lediglich fünf diskrete Intensitätsstufen (vgl. dazu Abschn. 5.1.2).

5.1.4. Abstandsmaße

Im Abschnitt 5.1.1 zur Klassifikation wurde bereits auf die ausführungsbezogenen Unterschiede und Ähnlichkeiten zwischen den zwölf Fazialisübungen eingegangen. Die Beschreibung der Ähnlichkeiten bezog sich dabei im Wesentlichen auf die Translation der Landmarken und die Aktivierung bestimmter Muskelgruppen. Im Folgenden wird diese eher intuitive Herangehensweise weiterentwickelt. Dabei wird beschrieben, wie sich Ähnlichkeiten auf Observationsebene quantifizieren lassen und in welcher Weise sie als Bewertungsmaß für die Qualität der Übungsausführung eingesetzt werden kön-

(a) (b) (c) (d) (e)

Abbildung 5.5.: **(a)** Für die dargestellte physische Mobilisierungsübung (Anheben des Arms von $\alpha = 0°$ bis $\alpha = 90°$) lassen sich Anfangs- und Endzustände definieren. Sie können in die Bewertung der Übungsdurchführung einfließen. Die komplexe Gesichtsmuskulatur erlaubt vielfältige, teilweise voneinander unabhängige, Bewegungen, deren Beschreibung trainiertes Personal erfordert (siehe dazu FACS in Abschn. 5.1.2). **(b)** Übung *Kuss*, ausgeführt durch eine gesunde Person. **(c)-(e)** Übung *Kuss*, ausgeführt durch Patienten mit Mimikdysfunktionen.

nen. Diese Vorgehensweise liegt nah, da die Aufgabe eines Feedbacksystems im Prinzip darin besteht, Un-/Ähnlichkeiten zwischen der Übungsausführung eines Patienten und einer Zielübung zu bestimmen.

Eng verknüpft mit den Ähnlichkeitsmaßen sind Distanzmaße. Beide werden im Folgenden unter dem Begriff der Abstandsmaße zusammengefasst. Bekannte Vertreter der Distanzmaße sind die euklidische Distanz und die Manhattan-Distanz [Duda et al., 2001]. Die euklidische Distanz $d(\mathbf{x}_A, \mathbf{x}_B)$ zwischen zwei Observationen A und B, repräsentiert durch die extrahierten Merkmalsvektoren $\mathbf{x}_A$ und $\mathbf{x}_B$, ergibt sich aus:

$$d(\mathbf{x}_A, \mathbf{x}_B) = \|\mathbf{x}_A - \mathbf{x}_B\|. \tag{5.1}$$

Ähnlichkeiten zwischen Observationen lassen sich beispielsweise aus einem trainierten Random-Forest (RF) ableiten [Cutler et al., 2012]. Nähere Informationen dazu sind im Abschnitt 5.2.2 beschrieben.

Im Folgenden werden euklidische Distanzen und Ähnlichkeiten hinsichtlich ihrer Eignung für das geplante Feedbackszenario untersucht und gegenübergestellt. Beide Abstandsmaße werden in der Regel paarweise ermittelt. Infolgedessen lassen sich die paarweisen Abstände von N Observationen in einer symmetrischen, $N \times N$ großen Abstandsmatrix $\mathbf{M}$ zusammenfassen [Cutler et al., 2012]. Im Fall der euklidischen Distanzmatrix entsprechen die Diagonaleinträge dem Wert 0, im Fall der RF-Ähnlichkeitsmatrix dem Wert 1. Beide Matrizen sind in der Regel ineinander überführbar. Die visuelle Interpretation beider Matrizen wird vereinfacht, indem die

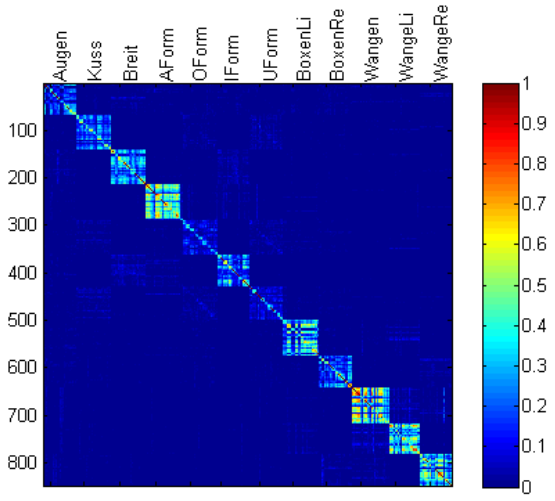


Abbildung 5.6.: Beispiel einer symmetrischen Ähnlichkeitsmatrix der Größe 847×847. Die Einträge sind so sortiert, dass Observationen gleicher Klassen nebeneinander liegen. Dadurch entstehen zusammenhängende Blöcke entlang der Hauptdiagonalen. Die Werte auf der Hauptdiagonalen selbst sind gleich 1, was per Definition dem höchstmöglichen Ähnlichkeitswert entspricht. Die dargestellten Ähnlichkeiten wurden aus einem trainierten Random-Forest ermittelt. Weiterführende Informationen dazu finden sich in den Abschnitten 5.2.1 und 5.2.2.

Sortierung der Einträge so erfolgt, dass Observationen gleicher Klassen nebeneinander liegen (siehe dazu Abb. 5.6).

Eine weitere Möglichkeit zur Visualisierung der in einer Abstandsmatrix $\mathbf{M}$ beschriebenen Observationsbeziehungen bildet die sogenannte metrische multidimensionale Skalierung (MDS) ([Cutler et al., 2012], [Wickelmaier, 2003]). Ihr Ziel ist eine weitestgehend distanzerhaltende Abbildung von, in einem hochdimensionalen Merkmalsraum berechneten, Distanzen auf einen niedrigdimensionalen Merkmalsraum.

Im Folgenden werden drei verschiedene Abstandsmatrizen ermittelt, auf Basis der MDS in eine niedrigdimensionalere Konfiguration transformiert und in einem zweidimensionalen Koordinatensystem visualisiert und verglichen. Die MDS wurde dabei mit der von der Matlab-Plattform bereitgestellten Methode *mdsprox*[1] durchgeführt. Die Basis der Auswertungen bildet ein Datensatz, welcher die Aufnahmen von zehn zufällig gewählten (gesunden) Personen umfasst (siehe Abschn. 2.3.2). In der Summe sind dies $N = 847$ Observationen, aus deren Tiefendaten Merkmalsvektoren extrahiert und zur Berechnung der Abstandsmatrizen $\mathbf{M}_{T1}$, $\mathbf{M}_{T2}$ und $\mathbf{M}_{T3}$ eingesetzt werden. Die Ausgangsmerkmalsvektoren umfassen insgesamt 922 Dimensionen, die sich aus

[1] `http://de.mathworks.com/help/stats/compacttreebagger.mdsprox.html`, letzter Zugriff: 25.08.2015

Merkmalsvariablen der in Kapitel 4 vorgestellten fünf Merkmalstypen zusammensetzen[2].

Der erste Teil der Auswertung bezieht sich auf die Matrix $\mathbf{M}_{T1}$. Deren Einträge setzen sich aus euklidischen Distanzen zusammen, welche auf Basis des 922-dimensionalen Ausgangsvektors berechnet wurden. Die aus der MDS resultierende, distanzerhaltende Konfiguration ist in der Abbildung 5.7a in einem zweidimensionalen Koordinatensystem dargestellt, wobei dessen Achsen den Komponenten mit den größten Eigenwerten entsprechen. Die gezeigte Punktwolke erscheint weitestgehend strukturlos. Eine Anordnung der Observationen entsprechend der annotierten Übungsklassen ist in den dargestellten Distanzbeziehungen nicht erkennbar.

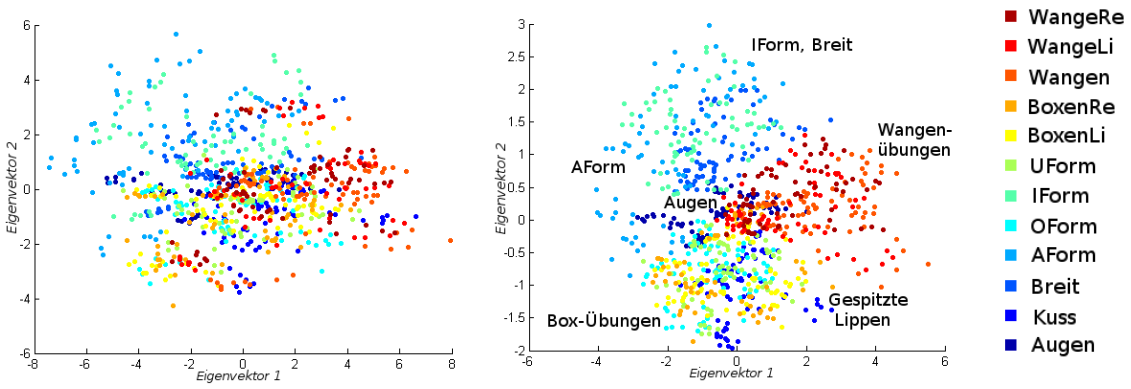
Die Vorgehensweise bei Analyse der Matrix $\mathbf{M}_{T2}$ gestaltete sich im Wesentlichen analog. Den Unterschied bildet eine vorgeschaltete Merkmalsselektion. Diese erfolgte grenzwertbasiert auf Basis der Mutual-Information (MI) und resultierte in einem reduzierten Merkmalsvektor, bestehend aus 221 Variablen [Steuer et al., 2002]. Wie die Abbildung 5.7b zeigt, führt die Beschränkung auf die 221 Merkmalsvariablen, welche die höchste MI zur ausgeführten Übung aufweisen, zu einer leicht verbesserten Repräsentation der annotierten Struktur. Trotz vereinzelt erkennbarer klassenspezifischer Häufungen, bleibt die Repräsentation der intra- und interklassen Distanzen unzufriedenstellend.

Ein wesentlicher Nachteil der euklidischen Distanzen besteht darin, dass bei ihrer Berechnung alle Merkmalsvariablen, unabhängig von ihrer Relevanz, gleichermaßen gewichtet werden. An dieser Stelle zeigt sich der Vorteil der aus einem Random-Forest abgeleiteten paarweisen Ähnlichkeiten. Bereits beim Erstellen des Random-Forests wird an jedem Knoten eines Entscheidungsbaumens eine Art Merkmalsselektion durchgeführt [Cutler et al., 2012]. Nähere Informationen zum Aufbau und Training der Random-Forests finden sich im Unterkapitel 5.2.

Im Rahmen der dritten Analyse soll daher untersucht werden, inwieweit die aus paarweisen Ähnlichkeiten bestehende Matrix $\mathbf{M}_{T3}$ die Struktur der Daten beschreibt. Der zu Grunde liegende Random-Forest umfasst 150 Entscheidungsbäume und wurde mit 922 Merkmalsvariablen trainiert. Die Ergebnisse der multidimensionale Skalierung von $\mathbf{M}_{T3}$ sind in den Abbildungen 5.8a und 5.8b gezeigt.

In beiden Streudiagrammen zeigt sich eine deutlich Gruppierung der Observationen hinsichtlich ihrer Klassenzugehörigkeit. Dies wird nicht zuletzt dadurch begünstigt, dass es sich bei der RF-internen Merkmalsselektion um ein überwachtes Verfahren

[2]Für die genaue Zusammensetzung der Punktsignaturen, sowie der Krümmungs- und HON-Merkmalsdeskriptoren siehe Tabelle 5.1. Die DWM-Merkmalsdeskriptoren wurden in reduzierter Form gemäß Gleichung 5.10 auf der Seite 155 extrahiert.



(a) (b) (c)

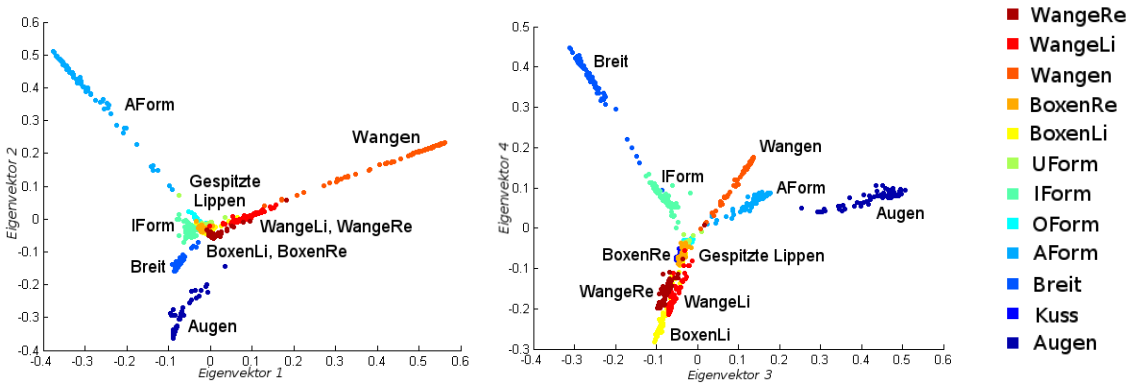
Abbildung 5.7.: Beispielergebnisse der MDS. Dargestellt sind die Hauptachsen mit den größten Eigenwerten (somit auch nur ein Ausschnitt der Distanzbeziehungen). Jeder der 847 Punkte entspricht einer Trainingsobservation und ist zum Zweck der Visualisierung entsprechend seiner Klassenzugehörigkeit eingefärbt, die MDS selbst ist jedoch ein unüberwachtes Analyseverfahren. **(a)** MDS-Ergebnisse unter Verwendung der euklidischen Distanzmatrix $\mathbf{M}_{T1}$. **(b)** MDS-Ergebnisse unter Verwendung der euklidischen Distanzmatrix $\mathbf{M}_{T2}$. **(c)** Legende für die Zuordnung der Farben zu den Klassen.

handelt, während die Berechnung der euklidischen Distanzen ohne Einbezug der Klassenzugehörigkeit erfolgt.

Neben den Intraklassen-Distanzen sind insbesondere die Interklassen-Distanzen aufschlussreich, da kein annotiertes Expertenwissen über mögliche Übungsklassenbeziehungen in die Erstellung des Random-Forest geflossen ist. In beiden Streudiagrammen ist eine ausgeprägte Nachbarschaft zwischen solchen Klassen erkennbar, die sich, aufgrund der Aktivierung derselben Muskelgruppen, auch visuell und anatomisch ähnlicher sind. Am deutlichsten zeigt sich dies bei der gehäuften Anordnung der Klassencluster der Übungen *Kuss*, *UForm* und *OForm* (gespitzte Lippen), denen allen eine Aktivierung des Mundringmuskels zu Grunde liegt (siehe dazu Anhang A.1). Ebenfalls in Nachbarschaft finden sich die Cluster der Fazialisübungen *Breit* und *IForm*, deren laterale Bewegung der Mundwinkel unter anderem durch eine Kontraktion des Musculus risorius sowie des kleinen Jochbeinmuskels erfolgt.

5.1.5. Zusammenfassung

Der Fokus dieses Unterkapitels lag auf der Identifikation von geeigneten technischen Verfahren für die Erstellung eines Mimiktraining-Feedbacksystems. Dazu wurde diskutiert, inwieweit sich klassifikations-, regressions- und abstandsmaßbasierte Ansätze



(a) (b) (c)

Abbildung 5.8.: Ergebnisse aus der Anwendung der MDS auf die Matrix $\mathbf{M}_{T3}$. Die Streudiagramme sind aufgespannt durch die ersten vier Eigenvektoren (E) mit den größten Eigenwerten und decken 23% der Varianz in den Daten ab. **(a)** E_1 und E_2. **(b)** E_3 und E_4. **(c)** Legende für die Zuordnung der Farben zu den Klassen.

zur Ableitung von Feedback aus den extrahierten Merkmalsdeskriptoren eignen. Desweiteren wurde auf das Facial-Action-Coding-System (FACS) eingegangen. Sowohl das FACS als auch die Regression stellen jedoch spezielle Anforderungen an die zu Grunde liegende Ground-Truth und werden daher im Rahmen dieser Arbeit nicht weiter verfolgt.

Im letzten Abschnitt wurden Abstandmaße, bestehend aus euklidischen Distanzen und paarweisen Ähnlichkeiten, vorgestellt. Die Aufgabe eines Feedbacksystems lässt sich im Prinzip als eine Art Abstandsbestimmung charakterisieren, bei welcher die Übungsausführung eines Patienten mit einer Zielübung verglichen und die Unterschiede in einer gewissen Form quantifiziert werden. Bei einer Gegenüberstellung der verschiedenen Abstandsmaße auf Basis einer multidimensionalen Skalierung waren die aus einem Random-Forest abgeleiteten paarweisen Ähnlichkeiten am besten dazu geeignet, die annotierte Struktur der Ground-Truth abzubilden. Sie sind daher der Ausgangspunkt für das in dieser Arbeit entwickelte Feedbackverfahren, welches im folgenden Unterkapitel vorgestellt wird.

5.2. Eigener Ansatz zur Feedbackerzeugung

Das vorliegende Unterkapitel beschreibt ein im Rahmen dieser Arbeit entwickeltes Verfahren zur Erzeugung von globalem und lokalem Feedback auf Basis extrahierter Merkmalsdeskriptoren. Die Ableitung des Feedbacks erfolgt (nahezu) unüberwacht

direkt aus den Trainingsdaten. Dies bedeutet, dass eine Annotation der Übungsklasse, nicht jedoch der Übungsqualität innerhalb der Ground-Truth, erforderlich ist. Ebenso muss die vom Patienten auszuführende Übung bekannt sein. Diese wird im Folgenden auch als Zielübung oder Vorgabeübung bezeichnet.

Das Unterkapitel umfasst zwei Themenschwerpunkte. Der erste Teil beschreibt die Grundlagen der Random-Forests und der daraus abzuleitenden paarweisen Ähnlichkeiten (Abschn. 5.2.1 und 5.2.2). Im zweiten Teil werden die entwickelten Methoden zur Erstellung von globalem und lokalem Feedback beschrieben (Abschn. 5.2.3 und 5.2.4).

5.2.1. Random-Forests

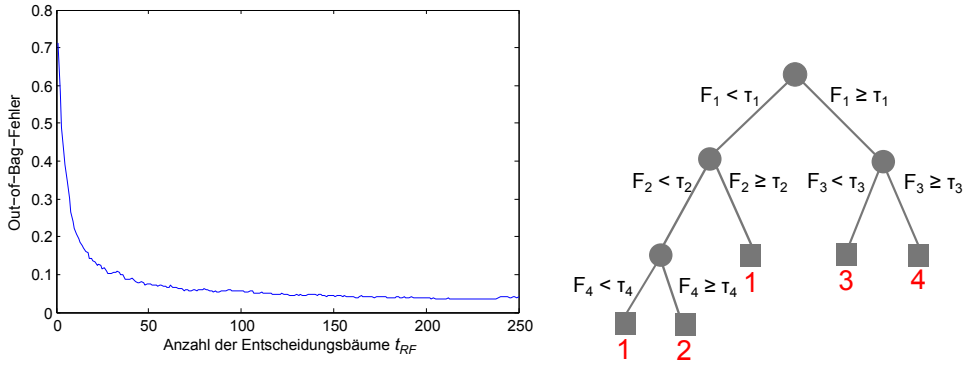
Als Random-Forests (RF) werden Ensembles aus mehreren Entscheidungsbäumen bezeichnet, die sowohl zur Klassifikation als auch zur Regression eingesetzt werden können [Cutler et al., 2012]. Im Fall der Klassifikation ergibt sich die resultierende Klasse aus einem Mehrheitsvoting der Einzelbaumergebnisse. Die entstandene Baumstruktur ermöglicht jedoch auch eine tiefergehende Analyse der Daten über die diskrete Klasse hinaus, wie beispielsweise die Bestimmung von Ähnlichkeiten zwischen beliebigen Observationspaaren. Beispiele für die Nutzung dieser Ähnlichkeiten wurden bereits in der Abbildung 5.8 gezeigt.

Zur Erstellung und Anwendung der Random-Forests wird die von Matlab bereitgestellte *TreeBagger*-Klasse[3] eingesetzt. Die gewählte Anzahl der Bäume pro RF beträgt in dieser Arbeit $t_{RF} = 150$. Im Allgemeinen ist der Einfluss dieses Parameters auf die Klassifikationsergebnisse gering, solange t_{RF} nicht zu klein gewählt wird.

Um eine Überanpassung an die Trainingsdaten zu vermeiden, wird jeder Baum auf einer Untermenge der Daten, dem sogenannten Bootstrap-Datensatz, trainiert. Dieser wird durch zufälliges Ziehen mit Zurücklegen aus allen Trainingsobservationen und für jeden Baum gesondert bestimmt. Die verbliebenen Observationen werden als Out-of-Bag-Daten (OOB-Daten) bezeichnet und können zur Berechnung von Bewertungsmaßen eingesetzt werden, für die sonst eine Kreuzvalidierung erforderlich wäre. Ein Beispiel hierfür ist der Out-of-Bag-Fehler (OOB-Fehler), welcher die, unter Einsatz der OOB-Daten ermittelte, mittlere Falschklassifikationsrate beschreibt. Die Abbildung 5.9a zeigt einen beispielhaften Verlauf des OOB-Fehlers in Abhängigkeit von t_{RF}.

Jeder Entscheidungsbaum besteht aus mehreren Kanten, Knoten und Endknoten, wie die vereinfachte Baumstruktur in Abbildung 5.9b verdeutlicht. Die Kanten werden auch als Zweige bezeichnet und die Endknoten als Blätter. Zu Beginn des Trai-

[3] `http://de.mathworks.com/help/stats/treebagger.html`, letzter Zugriff: 20.08.2015



(a) (b)

Abbildung 5.9.: **(a)** Verlauf des Out-of-Bag-Fehlers in Abhängigkeit von t_{RF} für einen zufällig gewählten Trainingsdatensatz, bestehend aus 10 Personen. Bei 150 Bäumen werden die OOB-Observationen im Mittel in 5,31 % der Fälle einer falschen Klasse zugeordnet (250 Bäume: 4,49 %). **(b)** Schematisches Beispiel eines einzelnen Entscheidungsbaumes. Knoten sind durch Kreise und Blätter durch Rechtecke visualisiert. Jedem Endknoten ist eine entsprechende Klasse zugeordnet, welche hier durch eine rote Ziffer symbolisiert wird. Testobservationen bekommen die Klasse des Endknotens zugewiesen, in welchem sie nach Durchlaufen des Entscheidungsbaumes landen.

nings wird der Bootstrap-Datensatz am Wurzelknoten anhand eines Grenzwertes τ_1 in zwei Gruppen aufgespalten, welche im Anschluss wiederum in zwei Gruppen unterteilt werden. Dieses Schema wird wiederholt, bis in jeder Gruppe ausschließlich die Observationen einer Klasse enthalten sind und somit einen Endknoten erreicht wurde. Die Grenzwerte τ_x ergeben sich dabei in zwei Schritten. Zuerst wird für den entsprechenden Knoten zufallsbasiert eine Untermenge von $d = \sqrt{D}$ Merkmalsvariablen gezogen, wobei D der Anzahl aller Merkmalsdimensionen entspricht. Danach erfolgt, ausgehend von der Untermenge d, eine Entscheidung für das Merkmal, welches die beste grenzwertbasierte Trennung der Daten erzielt. Die Güte der Trennung wird über den sogenannten Gini-Index ermittelt [Cutler et al., 2012]. Nach dem Training des Random-Forest durchläuft eine zu klassifizierende Testobservation jeweils alle t_{RF} Entscheidungsbäume bis zu den Endknoten. Die Ergebnissklasse ergibt sich aus einem Mehrheitsvoting aller t_{RF} Endknotenklassen.

5.2.2. Ableitung paarweiser Ähnlichkeiten

Anhand der trainierten Entscheidungsbäume lassen sich die Nachbarschaftsbeziehungen (*engl.* proximities) zwischen zwei Observationen quantifizieren, wobei die Ob-

(a) (b) (c) (d)

Abbildung 5.10.: **(a)** Referenzobservation der Klasse *Kuss*. Die Merkmalsextraktion erfolgte auf den Tiefendaten, zur besseren Visualisierung werden an dieser Stelle jedoch die Farbabbildungen gezeigt. **(b)** Vergleichsobservation derselben Klasse. Mit der Referenzobservation wird ein Ähnlichkeitswert von 0,45 erzielt. Dies bedeutet, dass Referenz- und Vergleichsobservation in 45 % aller $t_{RF} = 150$ Bäume in einem gemeinsamen Blattknoten enden. **(c)** Die Vergleichsobservation der Übung *UForm* erzielt mit der Referenzobservation einen Ähnlichkeitswert von 0,23. **(d)** Der Vertreter der Wangenübung weist mit einer Ähnlichkeit von 0 den geringstmöglichen Wert auf. Dies bedeutet, dass Referenz- und Vergleichsobservationen in keinem der Entscheidungsbäume in einem gemeinsamen Blattknoten gelandet sind.

servationen auch unterschiedlichen Klassen zugeordnet sein können [Cutler et al., 2012]. Die Nachbarschaftsbeziehungen werden im Folgenden auch als paarweise Ähnlichkeiten bezeichnet. Um die Ähnlichkeit zwischen zwei Observationen zu bestimmen, werden diese zuerst durch alle t_{RF} Entscheidungsbäume geschickt. Anschließend wird die Anzahl m der gemeinsam geteilten Endknoten ermittelt. Der Ähnlichkeitswert ergibt sich aus dem Verhältnis m/t_{RF}. Anschauliche Beispielergebnisse finden sich in den Abbildungen 5.10a bis 5.10d. Sie verdeutlichen, dass die paarweisen Ähnlichkeiten Informationen über die Nachbarschaftsbeziehungen der Observationen im Merkmalsraum enthalten, die über die jeweilige Klassenzugehörigkeit hinaus geht. Weitere Informationen zu den paarweisen Ähnlichkeitsmaßen wurden bereits bei der Einführung der Abstandsmaße gegeben 5.1.4.

5.2.3. Globales Feedback

Im Folgenden wird die in dieser Arbeit entwickelte Methode zur Erzeugung von globalem Feedback vorgestellt. Das globale Feedback bezieht sich auf das ganze Gesicht, eine gesonderte Evaluation einzelner lokaler Regionen wird noch nicht durchgeführt. Details zur Ableitung von lokalem Feedback werden im Abschnitt 5.2.4 beschrieben.

Die auszuführende Übung k, mit $k \in \{1, ..., K\}$, ist durch einen Trainingsplan vorgegeben. Sie kann daher als bekannt vorausgesetzt und in die Bewertung einbezogen werden. Die grundlegende Idee des Verfahrens ist es, die Ähnlichkeiten zwischen der Testobservation und allen Trainingsobservationen zu bestimmen und diese zu den trainingsdateninternen Ähnlichkeiten ins Verhältnis zu setzen. Dabei werden ausschließlich die Trainingsdaten der auszuführenden Übung k betrachtet. Die Observationen der restlichen $K - 1$ Klassen sind lediglich für das Training des Random-Forest relevant.

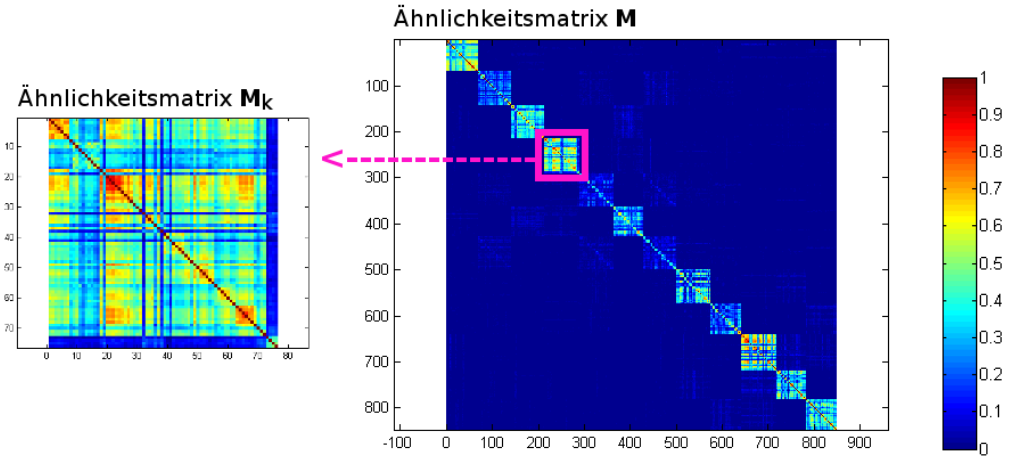
Die vorgestellte Methode stützt sich auf die Analyse eines statischen Einzelbildes. Die Auswertung kann jedoch in konstanten Zeitabständen wiederholt werden, um dynamisches Feedback zu erzeugen.

Der Algorithmus unterteilt sich im Wesentlichen in fünf Schritte. Die ersten beiden Schritte erfolgen offline und dienen der Erstellung von Modellen, welche während der Laufzeit Anwendung finden. Zu diesen Modellen zählen ein Random-Forest, sowie eine Ähnlichkeitsmatrix $\mathbf{M}$. Grundlegende theoretische und implementierungsbezogene Informationen zu den Random-Forests und der Ableitung der Ähnlichkeitswerte wurden bereits in den vorhergehenden Abschnitten 5.1.4, 5.2.1 und 5.2.2 beschrieben. Der Random-Forest wird auf Basis aller Trainingsobservationen, unabhängig von ihrer Klassenzugehörigkeit, ermittelt. Er umfasst in dieser Arbeit $t_{RF} = 150$ Entscheidungsbäume und dient als Basis für die Bestimmung der Ähnlichkeitsmatrix $\mathbf{M}$, welche die trainingsdateninternen, paarweisen Ähnlichkeiten beinhaltet.

In den zur Laufzeit folgenden Schritten wird ausschließlich das Segment $\mathbf{M}_k$ der Ähnlichkeitsmatrix $\mathbf{M}$ betrachtet, welches sich auf die Trainingsdaten der aktuell auszuführenden Übung k bezieht. Letztere wird im Folgenden auch als Vorgabe- oder Zielübung bezeichnet. Das Matrixsegment einer beliebigen Klasse k ist im Folgenden definiert durch:

$$\mathbf{M}_k = \begin{bmatrix} s_{1,1} & \cdots & s_{1,r} & \cdots & s_{1,R_k} \\ \vdots & & \vdots & & \vdots \\ s_{q,1} & \cdots & s_{q,r} & \cdots & s_{q,R_k} \\ \vdots & & \vdots & & \vdots \\ s_{Q_k,1} & \cdots & s_{Q_k,r} & \cdots & s_{Q_k,R_k} \end{bmatrix}. \tag{5.2}$$

Dabei beschreibt $s_{q,r}$ die Ähnlichkeit zwischen der q-ten und der r-ten Trainingsobservation. Aufgrund der Matrixsymmetrie gilt $Q_k = R_k$, wobei beide der Anzahl der Trainingsobservationen für die Fazialisübung k entsprechen. An dieser Stelle soll darauf hingewiesen werden, dass die ausführliche Schreibweise R_k im Folgenden zu



(a) (b)

Abbildung 5.11.: **(a)** Beispiel einer symmetrischen Ähnlichkeitsmatrix für zehn zufällig gewählte Trainingspersonen und zwölf therapeutische Übungsklassen. Bei der Feedbackerstellung wird lediglich das Matrixsegment benötigt, welches sich auf die Trainingsdaten der aktuell auszuführenden Übung k bezieht. Die Einträge des hervorgehobenen Segments repräsentieren die paarweisen Ähnlichkeiten der *AForm*-Trainingsobservationen ($k = 4$). Die Sortierung der Observationen innerhalb der Matrix $\mathbf{M}$ ist klassenbezogen und entspricht der Reihenfolge *Augen*, *Kuss*, *Breit*, *AForm*, *OForm*, *IForm*, *UForm*, *BoxenLi*, *BoxenRe*, *Wangen*, *WangeLi*, *WangeRe*. **(b)** Legende zur Farbskalierung der Ähnlichkeitsmatrix.

Gunsten einer besseren Lesbarkeit durch die Kurzschreibweise R ersetzt wird und sich immer auf die gegenwärtig betrachtete Vorgabeübung bezieht.

Die Abbildung 5.11a zeigt ein Beispiel einer Ähnlichkeitsmatrix für die insgesamt 847 Observationen von zehn zufällig gewählten Trainingspersonen. Das Matrixsegment $\mathbf{M}_k$, mit $k = 4$, beinhaltet die paarweisen Ähnlichkeiten zwischen den Trainingsobservationen der Klasse *AForm*. Zum Zweck der besseren Lesbarkeit wird im Folgenden die stringbasierte Schreibweise der integerbasierten vorgezogen (z.B. $\mathbf{M}_{AForm}$ statt $\mathbf{M}_4$).

Zur Laufzeit werden zu jedem vorgegebenen Auswertungszeitpunkt t die festgelegten Merkmale aus dem Gesicht des Patienten extrahiert und an den Random-Forest übergeben um diesen zu durchlaufen. Auf Basis der erreichten Endknoten können anschließend die paarweisen Ähnlichkeiten $a_{k,t,r}$ zwischen der aktuellen Übungsdurchführung und den Trainingsobservationen der Vorgabeübung k bestimmt werden. Sie werden in Form eines Vektors zusammengefasst:

$$\mathbf{a}_{k,t} = \begin{bmatrix} a_{k,t,1} & \cdots & a_{k,t,r} & \cdots & a_{k,t,R} \end{bmatrix}. \tag{5.3}$$

Der Vektor $\mathbf{a}_{k,t}$ ähnelt im Wesentlichen einer Zeile der Matrix $\mathbf{M}_k$, mit dem Unterschied, dass an Stelle einer Trainingsobservation eine Test- bzw. Patientenobservation mit allen R Trainingsobservationen verglichen wird. Für die paarweisen Ähnlichkeitswerte gilt per Definition $a_{k,t,r} \in [0;1]$. Das Produkt $(100 \cdot a_{k,t,r})$ gibt den Prozentsatz der Entscheidungsbäume an, in welchen die aktuelle Testobservation und die r-te Trainingsobservation in einem gemeinsamen Endknoten gelandet sind.

Die Gesamtheit der Vektoreinträge von $\mathbf{a}_{k,t}$ gibt einen ersten Eindruck über die, auf Basis der extrahierten Merkmalsdeskriptoren geschätzte, Ähnlichkeit einer Patientenobservation zu den Trainingsobservationen. Zur Veranschaulichung werden Ähnlichkeitsvektoren für drei zufällig gewählte Testobservationen der Klassen *Kuss*, *Augen* und *AForm* ermittelt und anhand von Histogrammen dargestellt. Die Histogramme sind in den Abbildungen 5.12a bis 5.12c gezeigt. Bei den Testobservationen handelt es sich um repräsentative Beispiele ihrer annotierten Klasse, da sie durch den Random-Forest korrekt klassifiziert wurden. Den Histogrammen nach weisen die beiden Testobservationen der Übungen *Augen* und *AForm* insgesamt eine größere Ähnlichkeiten zu ihren klassenidentischen Trainingsobservationen auf als die Testobservation der Übung *Kuss*. Auf eine, relativ betrachtet, schlechtere Ausführung der Übung *Kuss* lässt sich jedoch allein aus diesen Ergebnissen noch nicht schließen, wie anhand der Histogramme der Ähnlichkeitsmatrizen $\mathbf{M}_{Kuss}$, $\mathbf{M}_{Augen}$ und $\mathbf{M}_{AForm}$ deutlich wird. Die Histogramme sind in den Abbildungen 5.12d bis 5.12f gezeigt. Da $\mathbf{M}_k$ symmetrisch ist und alle Diagonaleinträge den Wert 1 haben, ist es ausreichend, zur Histogrammbildung die $(R^2 - R)/2$ Einträge oberhalb oder unterhalb der Hauptdiagonalen zu betrachten.

Zwischen den Histogrammen der Ähnlichkeitsmatrizen und denen ihrer korrespondierenden Ähnlichkeitsvektoren sind gewisse Übereinstimmungen erkennbar. Insgesamt betrachtet treffen auch innerhalb der Trainingsdaten Observationen der Übung *Kuss* seltener in einem Endknoten aufeinander als Observationen der Übungen *Augen* oder *AForm*. Bezieht man die Erkenntnisse der bisherigen Evaluationsergebnisse und Diskussionen aus Kapitel 4 bzw. Abschnitt 5.1.1 ein, lässt sich das seltenere Aufeinandertreffen durch falsch-positive Vertauschungen mit *Kuss*-ähnlichen Übungsklassen, wie *UForm* oder *OForm*, erklären. Diese Vertauschungen können durch subjektspezifische Variationen hinsichtlich der Übungsausführungen begünstigt werden (siehe dazu Abb. 5.13).

Vor dem Hintergrund der beschriebenen Problemstellung wurde ein Bewertungsmaß gesucht, welches die übungsspezifischen paarweisen Ähnlichkeiten $s_{q,r}$ der Trainingsdaten in den Evaluationsprozess einbezieht. Im Detail sieht das entwickelte Verfahren für die Evaluation einer Patientenobservation wie folgt aus. Der Vektor $\mathbf{a}_{k,t}$ wird, wie zuvor beschrieben, ermittelt und anschließend elementweise durchlaufen. Jedes Element

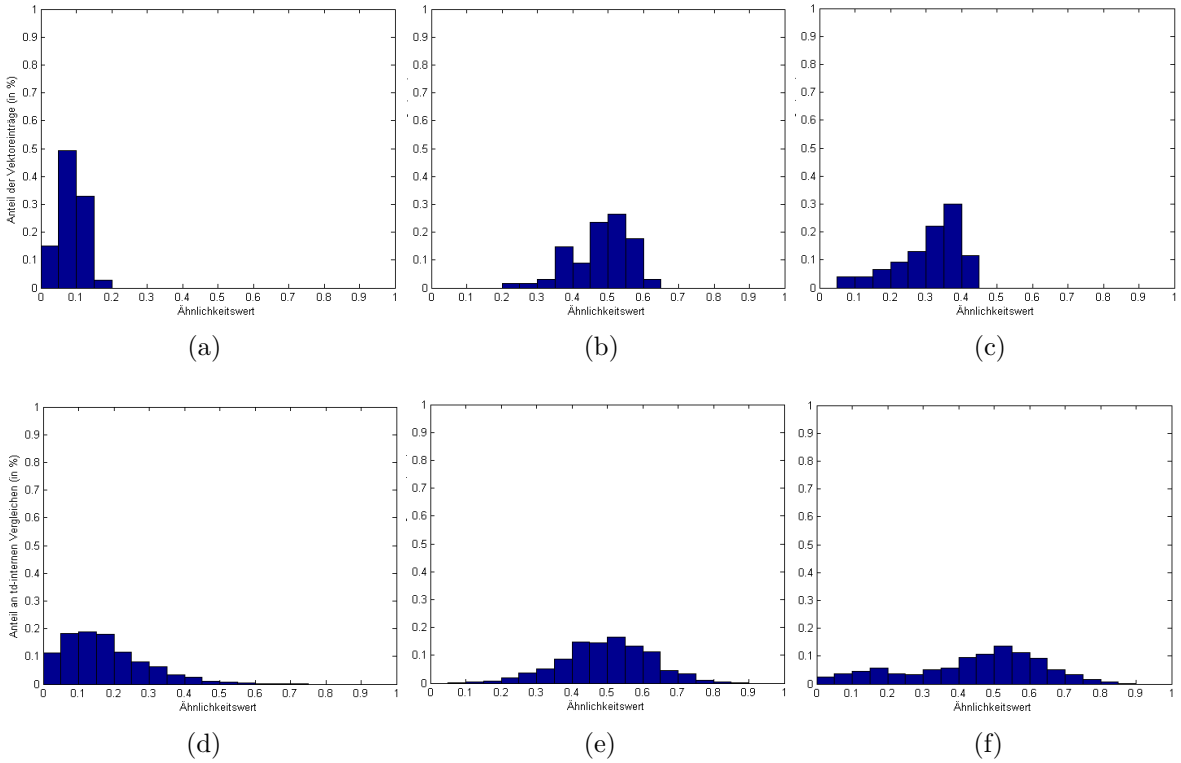


Abbildung 5.12.: **(a)**–**(c)** Histogramme über die Ähnlichkeitsvektoren $\mathbf{a}_{k,t}$ von drei zufällig gewählten Testobservationen der Klassen *Kuss*, *Augen* und *AForm* (v.l.n.r.). **(d)**-**(f)** Histogramme über $(R^2-R)/2$ Einträge der Ähnlichkeitsmatrizen M_k (mit k v.l.n.r. *Kuss*, *Augen* und *AForm*). Die Binbreite beträgt in allen Fällen 0,05.

$a_{k,t,r}$ entspricht bekannterweise der paarweisen Ähnlichkeit zwischen der Patientenobservation und der r-ten klassenidentischen Trainingsobservation. Diese paarweise Ähnlichkeit wird in Relation gesetzt zu den korrespondierenden paarweisen Ähnlichkeiten der Trainingsdaten. Korrespondierend sind per Definition alle $(Q-1)$ paarweisen Ähnlichkeiten zwischen der q-ten und der r-ten Trainingsobservation, wobei gilt $q \in \{1, ..., Q\}$ und $q \neq r$ (vgl. dazu Gleichung 5.2). Die Diagonalwerte der Matrix $\mathbf{M}_k$ mit $q = r$ werden nicht einbezogen, da sie jede Trainingsobservation mit sich selbst vergleichen und ihr Wert demzufolge gleich 1 ist. Die übrigen paarweisen Ähnlichkeiten der r-ten Spalte werden nun dahingehend ausgewertet, ob ihr Wert größer oder kleiner ist als $a_{k,t,r}$:

$$H(a_{k,t,r}, s_{q,r}) = \begin{cases} 1, & \text{falls } s_{q,r} < a_{k,t,r} \\ 0 & \text{sonst.} \end{cases} \quad \text{, es gilt } q \neq r. \tag{5.4}$$

Die ermittelten Werte für $H(a_{k,t,r}, s_{q,r})$, werden über alle $(Q-1)$ Spaltenelemente

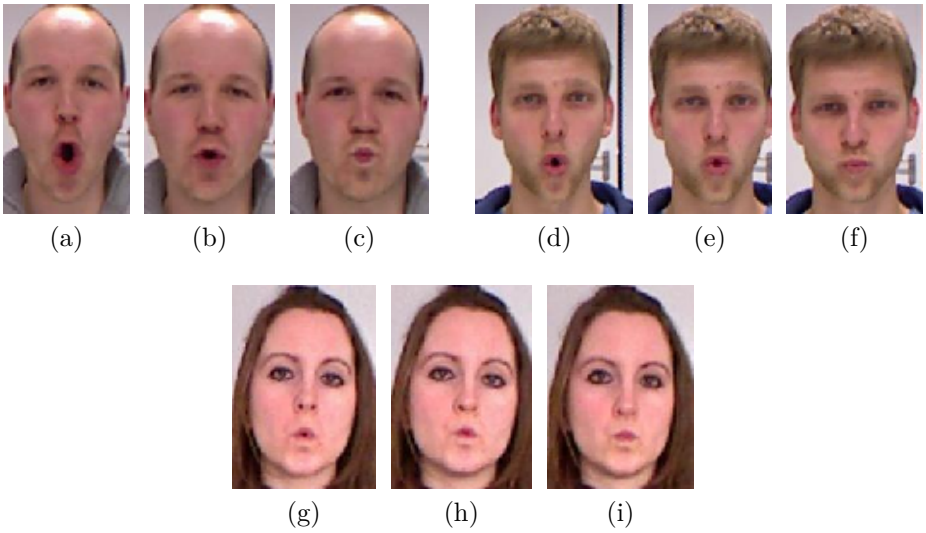


Abbildung 5.13.: **(a)**–**(i)** Personenspezifische Ausführungen der Mimikübungen *OForm*, *UForm* und *Kuss* bei gesunden Personen (Sortierung jeweils v.l.n.r.). Innerhalb eines Subjektes zeigen sich Regelmäßigkeiten, bspw. nimmt der Öffnungsgrad des Mundes von links nach rechts ab. Subjektübergreifend können sich die Ausführungen verschiedener Übungen, bezogen auf dieses Merkmal, ähneln (vgl. (b) *UForm* und (d) *OForm*).

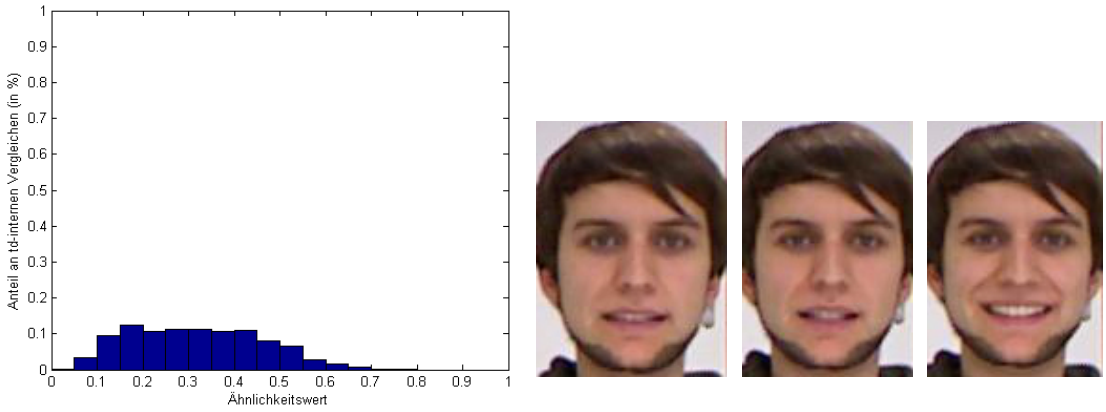
hinweg zu einem Skalar $v_{k,t,r}$ aufsummiert und gemittelt:

$$v_{k,t,r} = \frac{1}{(Q-1)} \sum_{q=1}^{Q} H(a_{k,t,r}, s_{q,r}) \text{ , es gilt } q \neq r. \tag{5.5}$$

Dabei kennzeichnet $v_{k,t,r}$ den prozentualen Anteil aller Trainingsobservationen, die seltener in einem gemeinsamen Endknoten mit der r-ten Trainingsobservation anzutreffen waren, als die Patientenobservation. Die in den Gleichungen 5.4 und 5.5 beschriebenen Schritte werden für alle R Einträge des Vektors $\mathbf{a}_{k,t}$ wiederholt und ihre Ergebnisse in einem Vektor $\mathbf{v}_{k,t}$ zusammengefasst:

$$\mathbf{v}_{k,t} = \begin{bmatrix} v_{k,t,1} & \cdots & v_{k,t,r} & \cdots & v_{k,t,R} \end{bmatrix}. \tag{5.6}$$

Die Gesamtheit der Vektoreinträge von $\mathbf{v}_{k,t}$ setzt somit die paarweisen Ähnlichkeiten zwischen der Testobservation und allen R Trainingsdaten in Relation zu den korrespondierenden paarweisen Ähnlichkeiten innerhalb der Trainingsdaten. Um ein skalares Bewertungsmaß zu erhalten, werden die Vektoreinträge zu einem Mittelwert zusammengefasst. Dies erfolgt anhand des Medians, um den Einfluss etwaiger Ausreißer zu minimieren:



(a) (b) (c) (d)

Abbildung 5.14.: **(a)** Histogramm über $(R^2 - R)/2$ Einträge der Ähnlichkeitsmatrix $\mathbf{M}_{IForm}$. Die Matrix basiert auf den Daten von zehn zufällig gewählten Trainingspersonen. **(b)**–**(d)** Durchführung der Zielübung *IForm* durch die Testperson. Das durch das Verfahren zugewiesene Feedback beträgt: (b) $\tilde{a} = 0{,}067$; $\tilde{v} = 0$ (c) $\tilde{a} = 0{,}07$; $\tilde{v} = 0$ (d) $\tilde{a} = 0{,}11$; $\tilde{v} = 0{,}01$.

$$\tilde{v}_{k,t} = \text{median}(\mathbf{v}_{k,t}). \tag{5.7}$$

Das resultierende Bewertungsmaß ist kontinuierlich, es gilt $\tilde{v}_{k,t} \in [0; 1]$. Ergänzend dazu wird der Median aller paarweisen Ähnlichkeiten $a_{k,t,r}$ ermittelt:

$$\tilde{a}_{k,t} = \text{median}(\mathbf{a}_{k,t}). \tag{5.8}$$

Beide Bewertungsmaße beziehen sich immer auf eine bestimmte Vorgabeübung k und einen Auswertungszeitpunkt t. Aus Gründen der Übersichtlichkeit werden im Folgenden die reduzierten Schreibweisen $\tilde{v}$ und $\tilde{a}$ verwendet. Die Verknüpfung der kontinuierlichen Bewertungsmaße $\tilde{v}$ und $\tilde{a}$ zu leichter erfassbarem diskretem Feedback ist in Abschnitt 5.4.2 beschrieben. In der Abbildung 5.14 werden drei Feedbackbeispiele für die Zielübung *IForm* vorgestellt, ergänzt um ein Histogramm über die paarweisen Ähnlichkeiten der Trainingsdaten. Die Histogramme über die jeweiligen Vektoren $\mathbf{a}_{k,t}$ und $\mathbf{v}_{k,t}$ der drei Testobservationen sind in der Abbildung 5.15 visualisiert.

5.2.4. Lokales Feedback

Für die Bestimmung des globalen Feedbacks wurde das Gesicht als Gesamtheit evaluiert. Eine örtliche Lokalisierung der Ausführungsfehler war nicht möglich. Um dem Patienten eine gezielte Korrektur seiner Übungsausführung zu ermöglichen, wird der

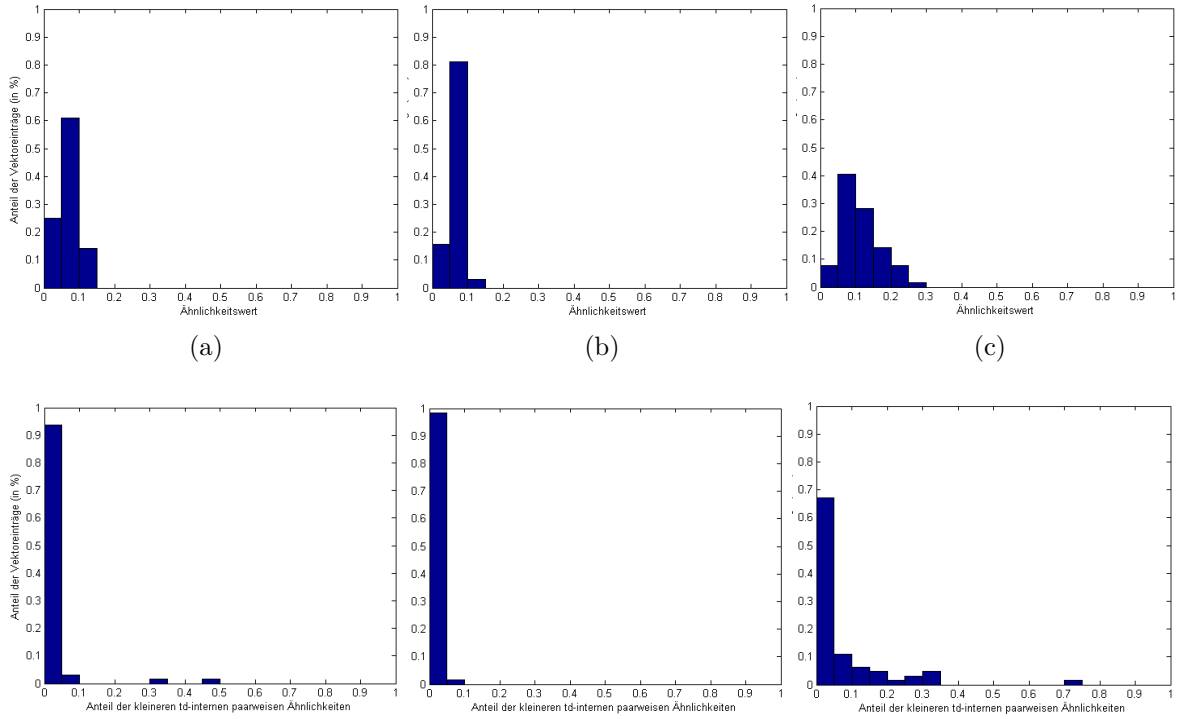


Abbildung 5.15.: **(a)**–**(c)** Histogramme über die Ähnlichkeitsvektoren $\mathbf{a}_{k,t}$ der in den Abbildungen 5.14b bis 5.14d gezeigten Testobservationen (Sortierung erneut v.l.n.r). **(d)**–**(f)** Histogramme über die resultierenden Vektoren $\mathbf{v}_{k,t}$.

vorgestellte Ansatz um eine lokale Feedbackmethode ergänzt. Diese entspricht in ihrem Ablauf der globalen Feedbackerzeugung, die Extraktion der Merkmale ist jedoch auf das zu evaluierende lokale Areal beschränkt. Folglich ist für jede Region ein gesonderter Random-Forest zu trainieren. Analog zu der im Abschnitt 5.2.3 beschriebenen Vorgehensweise, werden im Anschluss für jede Region s zwei Bewertungsmaße $\tilde{a}_s$ und $\tilde{v}_s$ ermittelt. Diese können nach Tabelle 5.4 regelbasiert zu einer diskreten Feedbackstufe verknüpft werden. Detaillierte experimentelle Auswertungen des globalen und lokalen Feedbacks finden sich in den Unterkapiteln 5.3 und 5.5.

5.3. Experimente zur Feedbackableitung

In den folgenden Abschnitten werden die grundlegenden technischen Bestandteile des Feedbackverfahrens szenariobezogen evaluiert. Der Fokus liegt dabei auf den Random-Forests und den paarweisen Ähnlichkeiten, welche das Fundament der entwickelten globalen und lokalen Feedbackerzeugung darstellen und in einem allgemeinen Kontext

in [Cutler et al., 2012] beschrieben wurden. Der Beitrag der folgenden Experimente umfasst jedoch weitere Aspekte:

- In Erweiterungen zu den Einzelevaluationen aus Kapitel 4 wird die Konkatenierung von Deskriptoren unterschiedlicher Merkmalstypen untersucht.
- Sowohl die globale als auch die lokale Klassifikation und Bestimmung der paarweisen Ähnlichkeiten werden evaluiert und können so verglichen werden.

Der Fokus der folgenden Analysen liegt somit auf verfahrensbezogenen Aspekten. Eine Evaluation des resultierenden Feedbacks findet sich im Unterkapitel 5.5.2. Den Experimentalabschnitten 5.3.2 und 5.3.3 ist eine einleitende Übersicht zu den Testbedingungen vorangestellt.

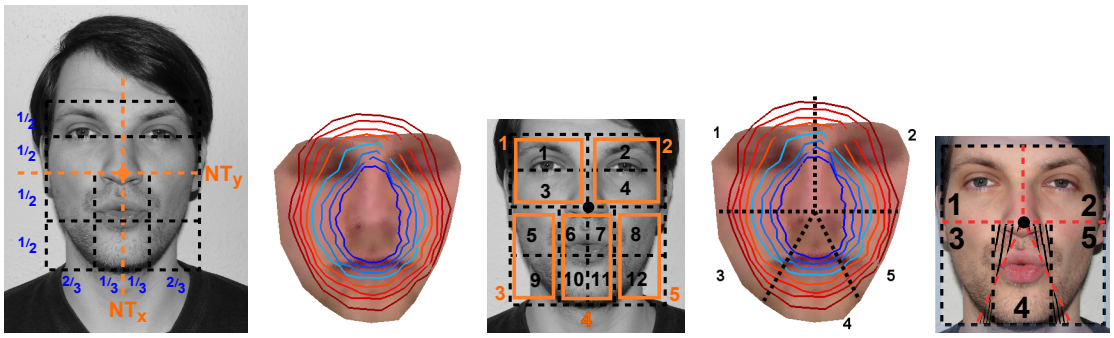
5.3.1. Testszenario

In Kapitel 4 wurden, in Vorbereitung zu den folgenden Auswertungen, verschiedene Verfahren zur Tiefenmerkmalsextraktion einzeln vorgestellt und evaluiert. Sie bilden die Basis der Feedbackerzeugung und werden in diesem Unterkapitel in kombinierter Form ausgewertet. Der grundsätzliche Aufbau der Experimente bleibt jedoch bestehen und wurde bereits im Unterkapitel 4.2 zusammenfassend beschrieben. Notwendige Anpassungen, die insbesondere durch die lokale Feedbackerzeugung entstehen, werden in diesem Abschnitt geordnet nach einzelnen Themenaspekten vorgestellt.

Extraktions- und Feedbackregionen

In dieser Arbeit wird zwischen sogenannten Extraktions- und Feedbackregionen unterschieden. Die Extraktionsregionen sind für die patchbasierte Merkmalsextraktion von Bedeutung. In den Unterkapiteln 4.5 und 4.6 konnte gezeigt werden, dass eine Erhöhung der Regionenanzahl bis zu einem gewissen Grad zu verbesserten mittleren Erkennungsraten führt. Aufbauend auf den erzielten Klassifikationsergebnissen wird für alle nachfolgenden Experimente eine Unterteilung des Gesichts in zwölf Extraktionsregionen festgelegt. Ausgangspunkt der Unterteilung ist die Bounding-Box, welche, ausgehend von der Position der Nasenspitze in vier Regionen unterteilt wird. Aus diesen Regionen ergeben sich, gemäß der in der Abbildung 5.16a definierten Verhältnisse, zwei beziehungsweise vier Subregionen. Für die Distanz-, Winkel- und Punktsignaturextraktion ist keine Definition von Extraktionsregionen erforderlich (siehe dazu Abb. 5.16b und 4.7).

Während die Extraktionsregionen somit im Wesentlichen für die Merkmalsextraktion von Bedeutung sind, grenzen die sogenannten Feedbackregionen die Areale ein, für



(a) (b) (c) (d) (e)

Abbildung 5.16.: **(a)** Geometrische Bestimmung der zwölf Extraktionsregionen aus der Bounding-Box des Gesichts. **(b)** Verlauf von acht Punktsignaturen unterschiedlicher Radien im Fall der globalen Merkmalsextraktion. **(c)** Zusammenfassung von zwei bzw. vier Extraktionsregionen zu insgesamt fünf Feedbackregionen. **(d)** Unterteilung der Punktsignaturen in fünf Segmente für die lokale Merkmalsextraktion. **(e)** Überscheidungen der rechteckigen und der radialen Feedbackregionen.

welche ortsbezogenes Feedback ermittelt und ausgegeben werden soll. In dieser Arbeit wurden insgesamt fünf solcher Areale festgelegt, bestehend aus der zentralen Mundregion, sowie den linken bzw. rechten Augen- und Wangenregionen. Da die Erzeugung des lokalen Feedbacks ausschließlich auf den extrahierten Merkmalsdeskriptoren einer lokalen Region basiert, ist die Zuweisung der einzelnen Merkmalsdimensionen zu den lokalen Regionen erforderlich. Im Fall der patchbasierten Merkmalsextraktion ist dies in einfacher Weise möglich, da sich eine Feedbackregion per Definition aus zwei oder vier Extraktionsregionen zusammensetzt. Die Abbildung 5.16c verdeutlich dies.

Die Zuordnung der radialen Punktsignaturen gestaltet sich komplizierter, da eine Signatur mehrere der vorgestellten Feedbackregionen durchläuft. Sie werden daher, wie in der Abbildung 5.16d gezeigt, in fünf Segmente unterteilt. Jeder Feedbackregion werden die Merkmalsvariablen zugeteilt, deren Extraktionsort in das entsprechende Segment fällt (vgl. dazu Abb. 4.15a bis 4.15c). Bei der Augenregion gelingt diese Zuordnung eindeutig. Im Fall der unteren Gesichtshälfte ergeben sich jedoch Überschneidungen, wie die Abbildung 5.16e veranschaulicht.

Für eine kleine Untermenge der 43 Distanz- und Winkelmerkmalsvariablen ist ebenfalls keine eindeutige Zuordnung zu einzelnen-Feedbackregion möglich (siehe dazu Abb. 4.7 und Tab. 4.3). Zu dieser Untermenge zählt unter anderem die Merkmalsvariable δ_{13}, welche die Beschaffenheit der Oberlippe in Form ihrer Länge quantifiziert. Bei Ausführung der Übung *Kuss* beschränkt sich die Position der Oberlippe im Wesentlichen auf die Mundregion 4, bei der Übung *Breit* erstreckt sie sich über alle drei

Merkmal	**Festlegung der merkmalsspezifischen Parameter**	**Dim.**
Punktsignaturen	Abtastintervall $\Delta\theta = 5{,}7°$ für 64 Dimensionen, Radien $r_i \in \{4\,\text{cm}; 4{,}5\,\text{cm}; 5\,\text{cm}; 5{,}5\,\text{cm}; 6\,\text{cm}; 6{,}5\,\text{cm}; 7\,\text{cm}; 7{,}5\,\text{cm}\}$	512
Krümmungen	12 Extraktionsregionen, 8 Bins pro Histogramm	96
HON	12 Extraktionsregionen, 5×5-Bins pro bivariatem Histogramm	300
Distanzen	keine freien Parameter	16
Winkel	keine freien Parameter	27
Merkmalsvektor (global)	-	$\sum$: 951

Tabelle 5.1.: Festlegung der Parameterwerte für die einzelnen Merkmalstypen. Die dritte Spalte enthält die Dimensionsanzahl der merkmalsspezifischen Subvektoren für den Fall der globalen Klassifikation. Der konkatenierte Ergebnisvektor umfasst insgesamt 951 Dimensionen.

Feedbackregionen der unteren Gesichtshälfte. Infolgedessen sind einige der Distanz- und Winkelvariablen im Rahmen des Evaluationsprozesses mehreren lokalen Regionen zugeordnet. Eine Übersicht dazu findet sich in der Tabelle B.4 im Anhang.

Merkmalsextraktion

Im Unterschied zu den Experimenten in Kapitel 4, bei welchen die Einzelevaluationen der fünf Merkmalstypen im Mittelpunkt standen, werden im Folgenden Kombinationen aus diesen evaluiert. Die Belegung der merkmalsspezifischen Parameter stützt sich dabei auf die Ergebnisse der Einzelevaluationen. Eine Zusammenfassung der gewählten Belegungen ist in der Tabelle 5.1 aufgeführt. Notwendige Anpassungen für die lokale Merkmalsextraktion, beispielsweise in Bezug auf die Anzahl der Extraktionsregionen, wurden bereits im vorhergehenden Abschnitt thematisiert.

Bei der globalen Feedbackableitung resultiert die Konkatenierung der fünf Merkmalstypen in einen Merkmalsvektor mit 951 Einträgen. Die Längen der lokalen Merkmalsvektoren variieren in Abhängigkeit von der jeweiligen Feedbackregion (Region 1-5: 207, 207, 167, 227 und 167).

Klassifikation

Random-Forests und die aus ihnen abgeleiteten paarweisen Ähnlichkeiten bilden die wesentlichen Bestandteile der entwickelten Methode zur Erzeugung von globalem und lokalem Feedback. Aus diesem Grund soll untersucht werden, inwieweit Random-Forests dazu geeignet sind, die in der Ground-Truth enthaltene Struktur abzubil-

den und eine Testobservation der korrekten Übungsklasse zuzuordnen. Zur Einordnung der Ergebnisse werden zusätzlich lineare Multiklassen-Supportvektormaschinen (lineare SVMs) und k-Nearest-Neighbor-Klassifikatoren (kNN-Klassifikatoren) in die Untersuchungen einbezogen.

Im Vergleich zu den 951 globalen Merkmalsvariablen ist die Anzahl der Observationen mit $n = 931$ eher gering, weshalb die Gefahr einer Überanpassung des Klassifikatormodells an die Trainingsdaten besteht. Sowohl lineare SVMs als auch Random-Forests weisen jedoch keine beziehungsweise nur eine sehr geringe Anfälligkeit für eine Überanpassung auf ([Hsu et al., 2003], [Cutler et al., 2012]). Der kNN-Algorithmus hingegen ist für sehr hochdimensionale Klassifikationsprobleme nicht beziehungsweise nur eingeschränkt geeignet. Dies ist auf seine Abhängigkeit von Distanzmaßen, wie beispielsweise der euklidischen Distanz, zurückzuführen [Beyer et al., 1999]. Zu Vergleichszwecken wird der kNN-Klassifikator dennoch in die Evaluation einbezogen[4]. Um die Zahl der Merkmalsvariablen zu reduzieren, wurden ergänzend Selektions- und Transformationsverfahren (z.B. PCA, LDA, Mutual-Information) getestet. Dies führte jedoch zu keiner eindeutigen Verbesserung der Erkennungsraten, weshalb im Folgenden nicht näher darauf eingegangen wird.

5.3.2. Globale Klassifikation und Ähnlichkeitsbestimmung

Die Experimente in diesem Abschnitt stützen sich im Wesentlichen auf die automatisierte Klassifikation von zwölf Fazialisübungen. Für jeden der drei Klassifikatoren wurden sechs verschiedene Merkmalskonstellationen untersucht. Die resultierenden mittleren Erkennungsraten (MER) liegen zwischen 54,65 % und 84,73 %. Verglichen mit dem besten Ergebnis der Einzelmerkmalsevaluation (MER 75,40 %), welches für die Kombination von linearen SVMs und Punktsignaturen ermittelt wurde, wird durch die Konkatenierung von mehreren Merkmalen eine Verbesserung der Klassifikation erreicht (vgl. Abschn. 4.7.1).

Bei allen sechs Merkmalskonstellationen erzielten lineare SVMs die besten Ergebnisse, gefolgt von den Random-Forests. Der Einsatz von kNN-Klassifikatoren führte zu den, mit deutlichem Abstand, niedrigsten Erkennungsraten. Eine detaillierte Auflistung der Ergebnisse findet sich in der Tabelle 5.2. Zusätzlich sind alle mittleren Erkennungsraten in einem Säulendiagramm in der Abbildung 5.17 zusammengefasst,

[4]Die Bestimmung der Nachbaranzahl k erfolgt über eine 10-fache Kreuzvalidierung, wobei jede der zehn Teilmengen die Daten von genau einer der zehn Trainingspersonen umfasst. Auf diese Weise wird verhindert, dass eine Person sowohl Bestandteil des Trainings- als auch des Validierungsdatensatzes ist. Die Daten der verbliebenen elften Person werden anschließend zur Bestimmung der Erkennungsraten verwendet.

	alle M	**ohne HON**	**ohne DM**	**ohne KM**	**ohne WM**	**ohne PS**
lin. SVM	84,65 %	84,73 %	83,66 %	83,67 %	83,02 %	81,22 %
RF	81,03 %	80,41 %	82,19 %	78,43 %	76,32 %	79,09 %
kNN	65,87 %	70,68 %	64,78 %	59,40 %	58,70 %	54,65 %

Tabelle 5.2.: Detaillierte Auflistung der globalen mittleren Erkennungsraten für die verschiedenen Klassifikatoren und Merkmalskombinationen.

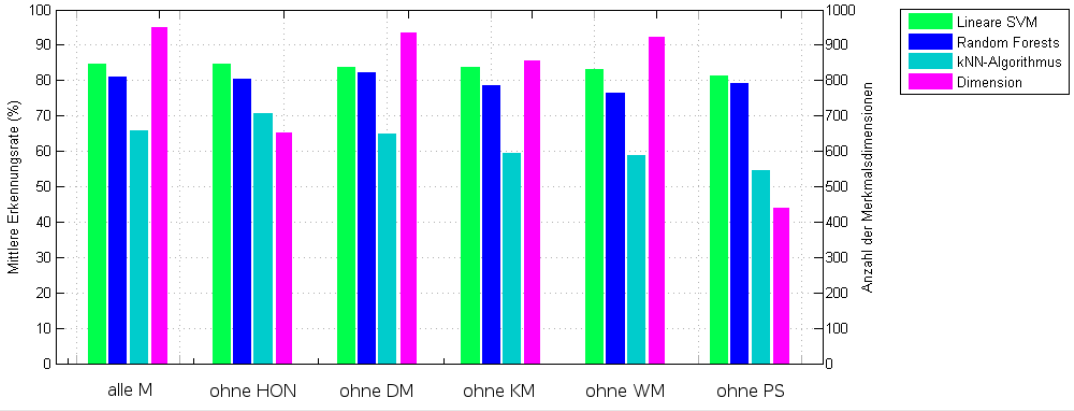
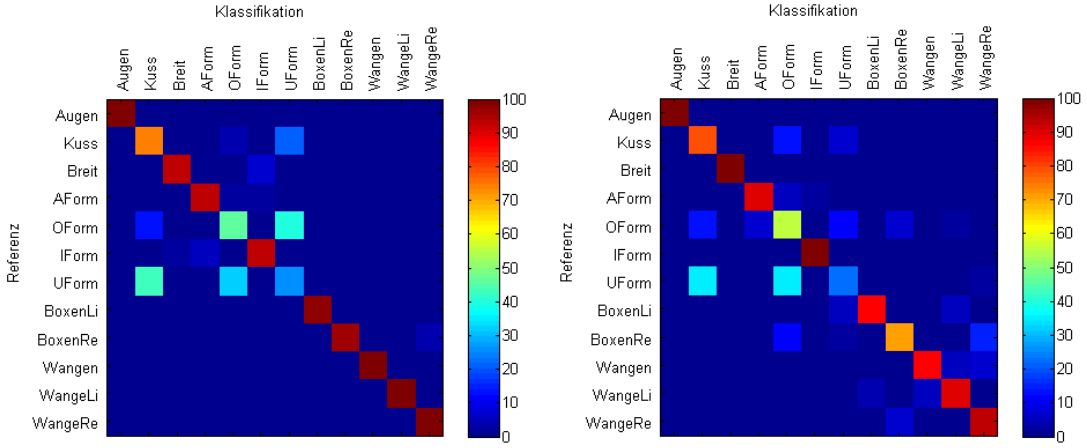


Abbildung 5.17.: Ergebnisse der globalen Klassifikation. Die blauen und grünen Säulen beziehen sich auf die linke Achse und stellen die erzielten mittleren Erkennungsraten unter Einsatz der verschiedenen Klassifikatoren dar. Die magentafarbenen Säulen beziehen sich auf die rechte Achse und bilden die Länge des extrahierten Merkmalsvektors ab. Abkürzungen: Merkmale (M), Histogramme orientierter Normalenvektoren (HON), Distanzmerkmale (DM), Krümmungsmerkmale (KM), Winkelmerkmale (WM), Punktsignaturen (PS).

jeweils ergänzt um die dazugehörige Länge des Merkmalsvektors.

Die erste der sechs Säulengruppen visualisiert die Ergebnisse für die Kombination von allen fünf Merkmalstypen (lin. SVM: 84,65 %, RF: 81,03 %, kNN: 65,87 %). Der konkatenierte, vollständige Merkmalsvektor umfasst, wie bereits erwähnt, 951 Dimensionen. Die übrigen Säulengruppen des Diagramms ergeben sich aus dem Ausschluss von einzelnen Merkmalstypen aus dem Klassifikationsprozess und resultieren dementsprechend in kürzeren Merkmalsvektoren.

Das Entfernen der HON- oder Distanzmerkmalsvariablen aus dem Klassifikationsprozess lässt bei den erzielten Erkennungsraten keine eindeutige Tendenz nach oben oder unten erkennen. Die Veränderungen im Vergleich zum Referenzverfahren mit vollständigem Merkmalssatz liegen mehrheitlich zwischen −1,09 und +1,16 Prozentpunkten. Eine Ausnahme bildet das Entfernen der HON-Merkmalsvariablen aus dem



(a) (b)

Abbildung 5.18.: Gemittelte Konfusionsmatrix aus den elf Konfusionsmatrizen einer 11-fachen, personenweise durchgeführten, Kreuzvalidierung. Der globale, 951-dimensionale Merkmalsvektor setzt sich gemäß Tabelle 5.1 zusammen. Verwendeter Klassifikator: **(a)** Lineare SVMs (MER: 84,65 %). **(b)** Random-Forests (MER: 81,03 %).

kNN-Klassifikationsprozess. Hierdurch kann eine Verbesserung um 4,81 Prozentpunkte erreicht werden. Gleichzeitig reduziert sich die Anzahl der Merkmalsvariablen deutlich auf 651.

Demgegenüber führt der Ausschluss der übrigen drei Merkmalstypen jeweils in allen Fällen zu Verschlechterungen bei den mittleren Erkennungsraten. Am größten fallen diese Rückgänge beim Einsatz des kNN-Klassifikators (−6,47 bis −11,22 Prozentpunkte) aus, am geringsten bei den linearen SVMs (−0,98 bis −3,43 Prozentpunkte). Random-Forests, welche in dieser Arbeit für die Berechnung der Ähnlichkeitswerte von wesentlicher Bedeutung sind, liegen im Mittelfeld (−1,94 bis −4,71 Prozentpunkte).

Die bisher beschriebenen mittleren Erkennungsraten stellen einen Mittelwert aus mehreren übungs- und personenübergreifenden Einzelraten dar. Die übungsspezifischen Erkennungsraten (ÜER) können aus den Konfusionsmatrizen ausgelesen werden, welche in den Abbildungen 5.18a und 5.18 visualisiert sind. Sowohl beim Einsatz der linearen SVMs als auch der Random-Forests zeigt sich, dass ein Mehrheit der Übungen vergleichsweise hohe Erkennungsraten aufweist und einige wenige Ausreißer zu einer deutlichen Reduzierung der mittleren Erkennungsrate führen.

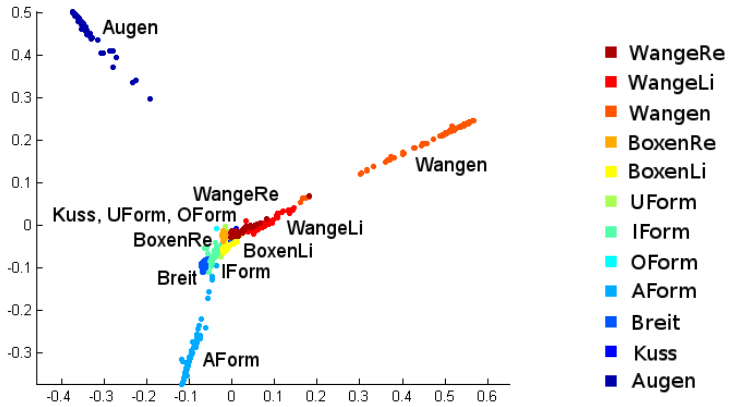
Insgesamt neun Diagonalwerte der sich aus der SVM-basierten Klassifikation ergebenden Konfusionsmatrix liegen zwischen 92,21 % und 100 %. Die mittlere Erkennungsrate beträgt, wie bereits erwähnt, 84,65 %. Die Ausreißer nach unten ergeben sich aus den Vertauschungen zwischen den Klassen *Kuss*, *UForm* und *OForm*, die bereits

bei den Einzelevaluationen der Merkmalstypen erkennbar waren (siehe Kap. 4). Die zuvor ebenfalls gehäuft aufgetretenen Vertauschungen *Breit-IForm* sind hingegen für den kombinierten Ansatz geringer ausgeprägt. Die Erkennungsrate der Übungsklasse *Kuss* ist mit 74,78 % vergleichsweise hoch, wohingegen die Übungsklassen *UForm* (ÜER: 25,22 %) und *OForm* (ÜER: 46,72 %) deutlich seltener korrekt zugeordnet werden. Da die Anzahl der Trainingsobservationen für alle Übungen weitestgehend ausbalanciert ist (vgl. Abschnitt 2.3.2), ist dies als Grund für die geringe ÜER der *UForm* auszuschließen. Wie anhand der Abbildungen 5.13a bis 5.13i deutlich wird, ist die Übung *UForm* eine Art Zwischenstadium der Ausführungen *Kuss* und *OForm*. Insgesamt werden 43,42 % bzw. 31,37 % der *UForm*-Testobservationen fälschlicherweise den Übungen *Kuss* bzw. *OForm* zugeordnet. Die Vertauschungen zwischen den Übungen *Kuss* und *OForm* sind mit 3,78 % und 12,99 % deutlich geringer ausgeprägt. Die vergleichsweise hohe ÜER der Klasse *Kuss* kann durch den vorgegebenen Mundschluss bedingt sein, der eine definierte Eigenschaft der entsprechenden Fazialisübung darstellt, während die Öffnung des Mundes bei den Übungen *UForm* und *OForm*, insbesondere personenabhängig, variiert.

Die Random-Forests, die für die in dieser Arbeit entwickelte Methode von besonderer Bedeutung sind, erzielen eine im Vergleich zu den linearen SVMs etwas geringere MER von 81,03 %. Dies spiegelt sich entsprechend in den ÜER wieder. Die niedrigsten Vertauschungsraten mit anderen Klassen ergeben sich für die Klassen *Augen*, *Breit* und *IForm*. Diese weisen daher ÜER zwischen 98,7 % und 100 % auf. Weitestgehend gleichbleibende Resultate ergeben sich für die Fazialisübungen *Kuss*, *UForm* und *OForm* (ÜER: 78,46 %, 23,38 %, 56,18 %). Verglichen mit den Ergebnissen der linearen SVMs, ist die Random-Forest-basierte Erkennung der wangenbezogenen Übungen weniger robust (ÜER: 70,78 % - 92,21 %). So liegen beispielsweise die Vertauschungsraten zwischen den Testobservationen der Übungen *BoxenRe* und *WangeRe* bei 14,29 % und 6,5 %.

Verglichen mit den Varianzen innerhalb der übungsspezifischen Erkennungsraten, sind die Schwankungen zwischen den personenspezifischen Erkennungsraten (PER) deutlich geringer ausgeprägt. Die personenspezifischen Erkennungsraten beschreiben den arithmetischen Mittelwert des prozentualen Anteils der korrekt erkannten Übungen, gesondert betrachtet für jede Testperson der 11-fachen personenweisen Kreuzvalidierung. Die aus dem Einsatz der linearen SVMs resultierenden PERs liegen zwischen 79,76 % und 89,35 % (arithm. Mittelwert: 84,65 %, Median: 85,00 %), die der Random-Forests zwischen 72,38 % und 85,71 % (arithm. Mittelwert: 81,03 %, Median: 80,71 %).

Neben den Konfusionsmatrizen stellen paarweise Ähnlichkeiten in Verbindung mit der multidimensionalen Skalierung (MDS) eine weitere Möglichkeit dar, um zu ana-



(a) (b)

Abbildung 5.19.: **(a)** Ergebnisse der multidimensionalen Skalierung auf Basis der Ähnlichkeitsmatrix von zehn zufällig gewählten Trainingspersonen und allen 951 globalen Merkmalsvariablen (alle fünf Merkmalstypen entsprechend Tab. 5.1). Dargestellt sind die ersten beiden Dimensionen der skalierten Koordinaten. Die Observationen sind entsprechend ihrer Klassenzugehörigkeit eingefärbt (siehe **(b)**).

lysieren inwieweit die extrahierten Merkmalswerte die annotierte Struktur der Daten repräsentieren. Sie bilden zudem das Fundament der in dieser Arbeit entwickelten Methode zur Feedbackerzeugung (siehe auch Abschn. 5.1.4 und 5.2.2). Während die Konfusionsmatrizen die Vertauschungen zwischen den Klassen beschreiben, erlauben die aus der MDS resultierenden skalierten Koordinaten Einblick in die Observationsbeziehungen.

Die Abbildung 5.19 visualisiert die Ähnlichkeitsbeziehungen von 847 Observationen von zehn zufällig bestimmten Personen. Alle 951 globalen Merkmalsvariablen wurden für das Training des Random-Forest und die Bestimmung der paarweisen Ähnlichkeiten verwendet. Dargestellt sind die beiden Dimensionen mit den größten Eigenwerten. Sie beschreiben 13 % der Varianz in den Daten. Jede Observation ist durch einen Punkt repräsentiert, welcher zur Veranschaulichung entsprechend seiner Klassenzugehörigkeit eingefärbt wurde. In der Karte ist eine klare Struktur erkennbar, Observationen mit derselben Klassenzugehörigkeit sind in weitestgehend zusammenhängenden Gruppen angeordnet. Ein weiteres Beispiel für einen reduzierten Merkmalssatz wurde bereits in der Abbildung 5.8 gezeigt.

5.3.3. Lokale Klassifikation und Ähnlichkeitsbestimmung

Wie in den Abschnitten 5.2.3 und 5.2.4 deutlich wurde, stimmen die Vorgehensweisen zur globalen und lokalen Feedbackerzeugung im Wesentlichen überein. Bei der

lokalen Methode ist lediglich eine Eingrenzung der Merkmalsextraktion auf die zu evaluierende lokale Region erforderlich. Im Rahmen der folgenden Experimente soll untersucht werden, welchen Einfluss eine Verkleinerung des Extraktionsareals auf die automatisierte Erkennung der zwölf Fazialisübungen und die Aussagekraft der paarweisen Ähnlichkeiten hat. Um die Random-Forest-basierten Klassifikationsergebnisse vergleichen und einordnen zu können, werden erneut lineare SVMs in die Experimente einbezogen. Die Evaluation mittels kNN-Klassifikatoren entfällt, da diese im vorhergehenden Abschnitt nur geringe Erkennungsraten erzielten.

Im Rahmen der folgenden Experimente wird für jede Region ein eigener Klassifikatormodell trainiert. Eingang in das Training finden dabei nur die Merkmalsvariablen, die aus der entsprechenden lokalen Region extrahiert wurden. Bei einer Vorgehensweise entsprechend Abschnitt 5.2.4 und 5.3.1, ergeben sich aus dem 951-dimensionalen, globalen Merkmalsvektor insgesamt fünf lokale Vektoren mit geringerer Variablenanzahl (Regionen 1-5: 207, 207, 167, 227, 167). In der Summe sind dies 975 Dimensionen, da, wie bereits im angeführten Abschnitt beschrieben, einige der Merkmalsvariablen mehreren Regionen zugeordnet wurden.

Wie die Abbildung 5.20 verdeutlicht, bewirkt die Eingrenzung des Extraktionsareales auf eine lokale Feedbackregion in allen Fällen einen Rückgang der mittleren Erkennungsraten. Während die globale mittlere Erkennungsrate beim Einsatz von linearen SVMs 84,65 % betrug, liegen die lokalen MERs (L-MER) zwischen 43,54 % und 78,62 %. Lokal trainierte Random-Forests erzielen L-MERs zwischen 39,26 % und 71,84 % (globale MER: 81,03 %). Eine detaillierte Auflistung aller Einzelergebnisse ist in der Tabelle 5.3 gegeben. In Abhängigkeit von der Feedbackregion, die der Merkmalsextraktion zu Grunde liegt, zeigen sich deutliche Unterschiede bei der automatisierten Übungserkennung. Sowohl unter Einsatz der linearen SVMs als auch der Random-Forests, werden für die Augenregionen die niedrigsten Erkennungsraten erzielt. Die Resultate für die Mundregion liegen im Mittelfeld. Bei einer Extraktion der Merkmale aus den Wangenregionen werden hingegen 70,92 % bis 78,62 % aller Testobservationen der korrekten Übungsklasse zugeordnet. Obwohl die aus den Wangenregionen extrahierten Merkmalsvektoren relativ gesehen die geringste Dimensionsanzahl (167) aufweisen, erreichen sie, in Verbindung mit linearen SVMs, Erkennungsraten, welche lediglich 6 bis 7 Prozentpunkte unterhalb der globalen MER liegen.

Zur Analyse der lokalen übungsspezifischen Erkennungsraten (L-ÜER) sind in den Abbildungen 5.21a bis 5.21c drei resultierende Konfusionsmatrizen der Random-Forest-basierten Klassifikation visualisiert. Sie beziehen sich auf die Augenregion 1, die Mundregion 4, sowie die Wangenregion 3. Die Lage der Feedbackregionen innerhalb des Gesichts wurde in der Abbildung 5.16c gezeigt.

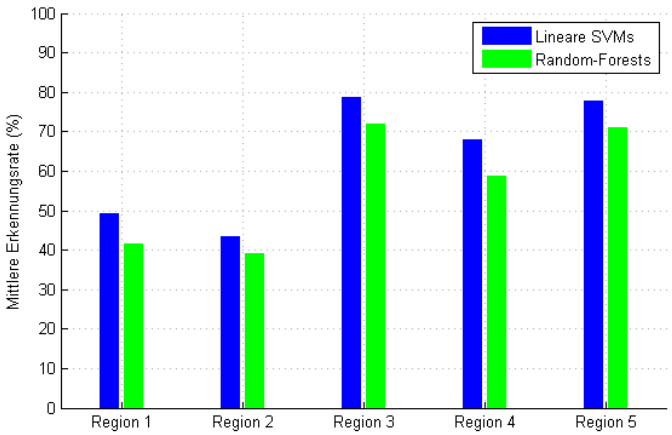


Abbildung 5.20.: Säulendiagramm der erzielten lokalen mittleren Erkennungsraten für die Klassifikation von zwölf Fazialisübungsklassen.

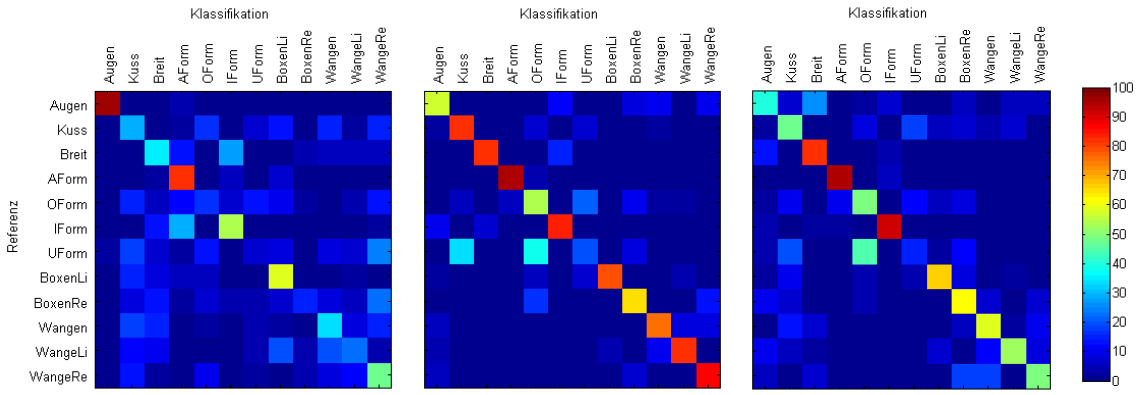
	Region 1	**Region 2**	**Region 3**	**Region 4**	**Region 5**
lin. SVM	49,18 %	43,54, %	78,62 %	67,95 %	77,85 %
RF	41,49 %	39,26 %	71,84 %	58,89 %	70,92 %

Tabelle 5.3.: Detaillierte Auflistung der lokalen mittleren Erkennungsraten für die Klassifikation mittels linearen SVMs und Random-Forests.

Verglichen mit der globalen Konfusionsmatrix in der Abbildung 5.18b, ist in allen lokalen Konfusionsmatrizen eine Zunahme der falsch-positiven Klassifikationsentscheidungen erkennbar. Am deutlichsten zeigen sich diese bei zu Grunde liegenden Merkmalsdeskriptoren, die aus der Augenregion extrahiert werden. Die Augenregion umfasst sowohl das Areal des Auges als auch das Areal der Wange auf Höhe des Jochbeins. Die lokale mittlere Erkennungsrate beträgt 41,49 %, wobei die höchste L-ÜER von 96,1 % für die Übungsklasse *Augen* erzielt wird. Der Lidschluss stellt ein markantes Merkmal der Übungsausführung *Augen* innerhalb der Augenregion 1 dar. Er wird durch die Merkmalsdeskriptoren δ_6, θ_9 und θ_{11} direkt extrahiert (vgl. Tab. 4.3).

Der Einsatz der extrahierten Mundmerkmalsdeskriptoren führt bei der Random-Forest-Klassifikation zu einer verbesserten L-MER von 58,89 %. Die höchsten L-ÜER werden für die Übungen *Breit*, *AForm* und *IForm* erzielt. Übungen, welche sich durch gespitzte Lippen oder einen stärkeren Wangenbezug auszeichnen, werden seltener korrekt zugeordnet.

Die höchste lokale MER wird bei der Klassifikation mittels Wangenmerkmalen erzielt (71,84 %, siehe auch Abb. 5.21b). Obwohl diese im vorgestellten Beispiel ausschließlich aus der, vom Patienten aus gesehen, rechten Wangenregion extrahiert wurden, erzielen auch die Übungklassen *BoxenLi* und *WangeLi* hohe Erkennungsraten.



(a) (b) (c) (d)

Abbildung 5.21.: Konfusionsmatrizen der Random-Forest-basierten Klassifikation für drei verschiedene Feedbackregionen. **(a)** Rechte Augenregion (L-MER: 41,49 %). **(b)** Rechte Wangenregion (L-MER: 71,84 %). **(c)** Mundregion (L-MER: 58,89 %). **(d)** Farbskala der Konfusionsmatrizen.

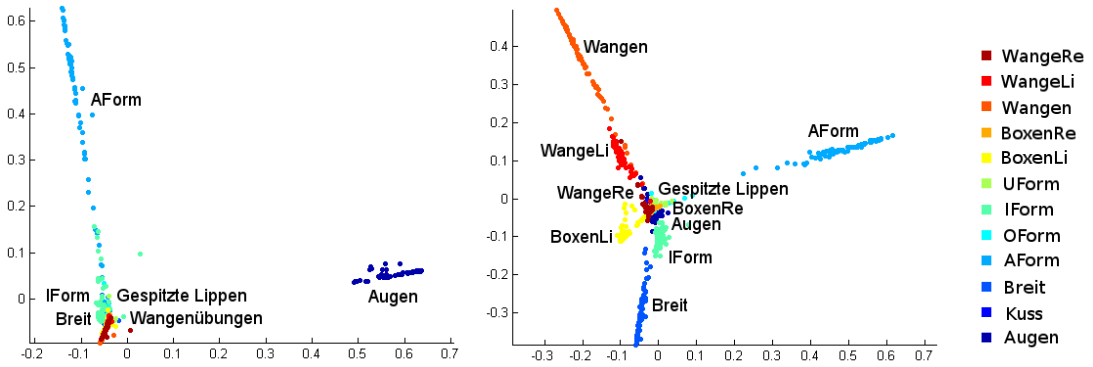
Desweiteren sind die Vertauschungsraten zwischen den Übungsklassen *Wangen* und *WangeRe* gering ausgeprägt, obwohl sich beide durch eine Wölbung der rechten Wange auszeichnen.

Die Ergebnisse der Konfusionsmatrizen spiegeln sich auch in den Ergebnissen der multidimensionalen Skalierung wider (siehe Abb. 5.22a und 5.22b). Zusammenfassend betrachtet, ist bei der lokalen Klassifikation Verbesserungsbedarf hinsichtlich der Diskriminanz der Merkmalsdeskriptoren erkennbar.

5.3.4. Zusammenfassung

In den beiden vorhergehenden Abschnitten wurden die zu Grunde liegenden Verfahren und Konzepte der lokalen und globalen Feedbackerzeugung evaluiert. In Erweiterung zu den Einzelevaluationen in Kapitel 4 wurde der kombinierte Einsatz der fünf Merkmalstypen zur Übungsklassifikation untersucht. Der Konkatenierung resultierte dabei in einer verbesserten globalen Erkennungsrate von bis zu 84,65 % (vorher: 75,40 %).

Während die globale Feedbackerzeugung auf den Merkmalen des ganzen Gesichts basiert, erfordert die lokale Feedbackerzeugung die Eingrenzung des Merkmalsextraktionsareals auf die zu bewertende lokale Region. Diese Eingrenzung führte zu einer Verringerung der mittleren Erkennungsraten auf 41,49 bis 71,84 Prozent. Hinsichtlich zukünftiger Arbeiten ist somit, insbesondere in Bezug auf die lokale Merkmalextraktion, eine Erweiterung und Verbesserung der Merkmalsextraktionsverfahren sinnvoll.



(a) (b) (c)

Abbildung 5.22.: Zweidimensionale, ähnlichkeitsrelationenbewahrende Abbildungen der paarweisen Ähnlichkeiten von zehn zufällig gewählten Trainingspersonen. **(a)** Augenregion 1 als Feedbackregion. **(b)** Wangenregion 3 als Feedbackregion. **(c)** Farblegende der zwölf Übungsklassen.

5.4. Implementierung des Prototypen

Der Fokus dieses Unterkapitels liegt auf der Umsetzung des Prototypen und der grafischen Visualisierung des Feedbacks. Die Einordnung dieser Themenpunkte in die technische Gesamtarchitektur wurde in der Abbildung 5.1 gezeigt. Der erste Abschnitt dieses Unterkapitels beinhaltet einleitende Vorbemerkungen. In den anschließenden Abschnitten werden die einzelnen Komponenten der grafischen Nutzeroberfläche vorgestellt und implementierungsbezogene Eckdaten des Prototypen aufgeführt.

5.4.1. Vorbemerkungen

In den Abschnitten 5.2.3 und 5.2.4 wurde die Ableitung der globalen und lokalen Bewertungsmaße aus den extrahierten Merkmalsdeskriptoren beschrieben. Die Bewertungsmaße liegen im Intervall $[0;1]$ und umfassen die Variablen $\tilde{a}$, $\tilde{v}$, $\tilde{a_s}$ und $\tilde{v_s}$. Der Index s definiert die lokale Feedbackregion, wobei gilt $s \in \{1,...,5\}$. Die Abbildung 5.23 enthält eine Übersicht über verschiedene Möglichkeiten zur Darstellung und Vermittlung dieses Feedbacks. Bei der Konzeption der Benutzerschnittstelle ist es sinnvoll, die physischen und kognitiven Fähigkeiten der angestrebten Zielgruppe einzubeziehen. Letztere kann sowohl (Vor-)Schulkinder mit myofunktioneller Dysfunktion als auch ältere Patienten im Rahmen der Schlaganfallrehabilitation umfassen. Eine eingehende Erörterung von Aspekten der Usability übersteigt jedoch Umfang und Zielstellung dieser Arbeit.

Zur Demonstration des Feedbackverfahrens und seiner Ergebnisse wurde für den

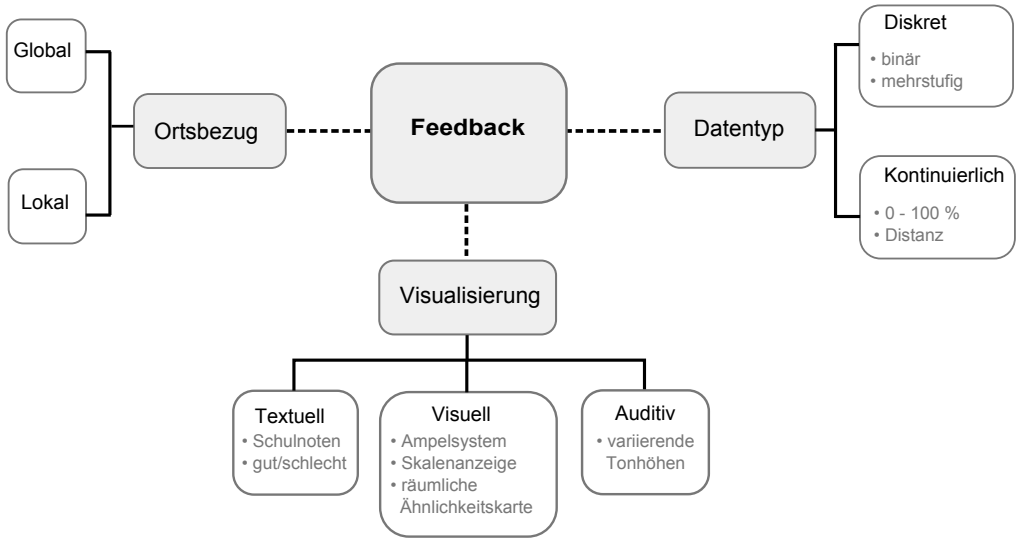


Abbildung 5.23.: Varianten und Eigenschaften des Feedbacks.

Prototypen eine visuelle Darstellung gewählt. Diese kann schnell erfasst werden und ermöglicht, im Gegensatz zur auditiven oder textuellen Rückmeldung, die übersichtliche parallele Darstellung verschiedener Informationen. Detailliertere Beschreibungen der einzelnen visuellen Feedbackkomponenten finden sich in den folgenden Abschnitten.

5.4.2. Aufbau der Prototyp-GUI

Die Nutzeroberfläche des implementierten Prototypen teilt sich im Wesentlichen in zwei Bereiche. Der erste Bereich dient der Anleitung des Patienten und zeigt die durch den Trainingsplan vorgegebene Übung an. Der zweite Bereich dient der Visualisierung des Feedbacks und umfasst insgesamt fünf verschiedene Feedbackkomponenten. Diese sind in der Abbildung 5.24a mit den Bezeichnern A bis E markiert und werden im Folgenden vorgestellt.

Feedbackkomponenten *A* und *B*

Die Komponenten A und B bilden die primären Feedbackelemente des implementierten Prototypen und dienen der Visualisierung des globalen Feedbacks. Die Komponente A umfasst eine Skalenanzeige, die das Bewertungsmaß $\tilde{v} \in [0; 1]$ visualisiert (Abb. 5.24b). Ein Skalenwert von $\tilde{v} = 0{,}4$ ist somit so zu interpretieren, dass im Median 40 % der korrespondierenden, trainingsdateninternen Observationsvergleiche geringere paarweise Ähnlichkeiten erzielen als die Testobservation zur r-ten Training-

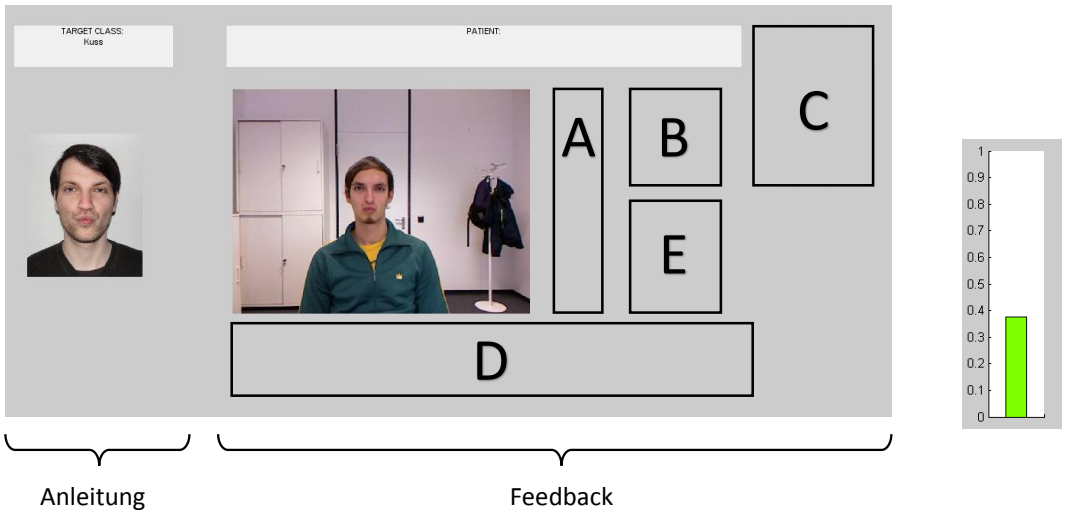


Abbildung 5.24.: **(a)** Anordnung der Elemente auf der Nutzeroberfläche. **(b)** Beispiel der Skalenanzeige der Komponente A.

sobservation. Es gilt $r = 1, ..., R$, wobei R der Anzahl der klassenidentischen Trainingsobservationen entspricht. Nähere Informationen zum Bewertungsmaß $\tilde{v}$ wurden im Abschnitt 5.2.3 beschrieben.

Der Vorteil des Bewertungsmaßes $\tilde{v}$ ist, dass es in komprimierter Form Informationen über die Ähnlichkeit der Testobservation zu den Trainingsdaten der auszuführenden Fazialisübung bereitstellt. Da sich die Trainingsdaten einer Übung jedoch ausschließlich aus korrekt ausgeführten Observationsbeispielen zusammensetzen, ist die Interpretation von $\tilde{v}$ als Skalenwert wenig intuitiv. So erreicht beispielsweise der Wert $\tilde{v} = 0{,}4$ nicht einmal die Hälfte der Skalenhöhe, entspricht jedoch, gemäß der Definition von $\tilde{v}$, einer Übungsausführung, deren extrahierte Merkmalsdeskriptoren im Wesentlichen denen der Trainingsdaten ähneln.

Aus diesem Grund wird die informative Skalenanzeige um die, vom Patienten intuitiver zu erfassende Komponente B, ergänzt. Diese besteht aus einem Smiley, dessen Gesichtsausdruck und Farbe der Übungsqualität entsprechend angepasst wird. Das Farbschema ist dabei an die Ampelfarben rot-orange-grün angelehnt. Für den Prototypen dieser Arbeit wurden insgesamt sechs verschiedene Smilies erstellt, die jeweils diskrete Bewertungsstufen repräsentieren. Die diskreten Bewertungsstufen werden dabei grenzwertbasiert aus den globalen Bewertungsmaßen $\tilde{v}$ und $\tilde{a}$ ermittelt. Eine Übersicht über die Grenzwerte ist in der Tabelle 5.4 gegeben. Die erste Stufe entspricht der niedrigsten Bewertung und tritt ein, wenn die Bedingung $\tilde{a} = 0$ erfüllt ist. In diesem Fall weisen mehr als 50% der R Test- und Trainingsobservationspaare eine paarweise Ähnlichkeit von 0 auf. Dies bedeutet, dass sie in keinem der t_{RF} Entscheidungsbäume

	Stufe 1	**Stufe 2**	**Stufe 3**	**Stufe 4**	**Stufe 5**	**Stufe 6**
						
B1	$\tilde{a} = 0$	$\tilde{a} > 0$	$\tilde{a} > 0$	$\tilde{a} > 0$	$\tilde{a} > 0$	$\tilde{a} > 0$
B2		$\tilde{v} = 0$	$\tilde{v} > 0$	$\tilde{v} > 0{,}05$	$\tilde{v} > 0{,}3$	$\tilde{v} > 0{,}5$
B3			$\tilde{v} \leq 0{,}05$	$\tilde{v} \leq 0{,}3$	$\tilde{v} \leq 0{,}5$	

Tabelle 5.4.: Sechs diskrete Feedbackstufen der Komponente B und die ihnen zugeordneten Symbole. Die Bestimmung der diskreten Stufe erfolgt grenzwertbasiert aus den kontinuierlichen Bewertungsmaßen $\tilde{v}$ und $\tilde{a}$. Alle Bedingungen einer Spalte müssen erfüllt sein ($B1 \wedge B2 \wedge B3$).

in einem gemeinsamen Endknoten gelandet sind. Die übrigen Feedbackstufen setzen ein $\tilde{a} > 0$ voraus und beziehen die trainingsdateninternen paarweisen Ähnlichkeiten der Vorgabeklasse ein (Details siehe Abschn. 5.2.3). Beispielergebnisse für das globale Feedback der GUI-Komponenten A und B werden im Abschnitt 5.5.2 evaluiert.

Feedbackkomponenten *C* und *D*

Die von den Komponenten A und B gegebenen Rückmeldungen an den Patienten bilden den wesentlichen Teil des (globalen) Feedbacks. Die ihnen zu Grunde liegenden Algorithmen stützen sich dabei nahezu ausschließlich auf die Trainingsobservationen, die der Klasse der auszuführenden Übung angehören. Die Motivation hinter dieser Vorgehensweise wurde bereits in den Abschnitten 5.1.1 und 5.1.4 deutlich. Eine Ausnahme hierbei bildet der Random-Forest (RF), welcher unter Einbeziehung der extrahierten Merkmalsdeskriptoren aller Trainingsobservationen trainiert wird.

Neben den paarweisen Ähnlichkeiten liefert der RF auch ein diskretes Klassifikationsergebnis für die Testobservation. Dieses Klassifikationsergebnis entspricht einer der zwölf Fazialisübungen. Die Komponente C zeigt Modellbild und Namen der zugeordneten Übung an. Das Klassifikationsergebnis des RF repräsentiert dabei die Klasse, welche der Testobservation nach Durchlaufen aller $t_{RF} = 150$ Entscheidungsbäume mehrheitlich zugeordnet wurde.

Die Auskunft der bisher beschriebenen Feedbackkomponenten bezieht sich lediglich auf eine Untermenge der Trainingsdaten. Das Feedback der Komponenten A und B basiert auf den paarweisen Ähnlichkeiten zwischen der Testobservation und den Trainingsobservationen der Vorgabeübung. Die Feedbackkomponente C benennt die durch den Klassifikator zugewiesene Übungsklasse. Dies kann die Vorgabeklasse oder eine

falsch-positive Klasse sein. Die Feedbackkomponente D ergänzt die grafische Benutzeroberfläche um eine Draufsicht, indem sie einen Überblick über die paarweisen Ähnlichkeiten zwischen der Testobservation und allen Trainingsobservationen, unabhängig von ihren Klassenzugehörigkeiten, bereitstellt. Aufbauend auf der Gleichung 5.3 lässt sich ein erweiterter Ähnlichkeitsvektor $\mathbf{a}_t$ definieren, der diese paarweisen Ähnlichkeiten umfasst:

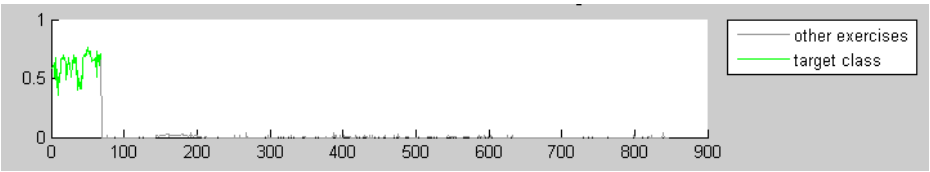
$$\mathbf{a}_t = \begin{bmatrix} \mathbf{a}_{1,t} & \cdots & \mathbf{a}_{k,t} & \cdots & \mathbf{a}_{K,t} \end{bmatrix}. \tag{5.9}$$

Der Subvektor $\mathbf{a}_{k,t}$, mit $k \in \{1, ..., K\}$, beinhaltet die paarweisen Ähnlichkeiten zwischen der Testobservation und den Trainingsobservationen der k-ten Klasse. Der Index t kennzeichnet den Auswertungszeitpunkt, auf den sich der Vektor $\mathbf{a}_t$ bezieht.

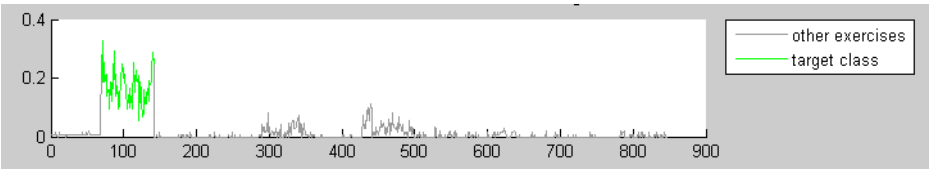
Für die Visualisierung auf der grafischen Benutzeroberfläche wird der Vektor in ein Schaubild geplottet. Auf der x-Achse ist der Index i der Trainingsobservation abgebildet, auf der y-Achse der Betrag der paarweisen Ähnlichkeit, die zwischen der Testobservation und der i-ten Trainingsobservation ermittelt wurde. Die Abbildungen 5.25a bis 5.25c zeigen drei beispielhafte Screenshots der Feedbackkomponente *D*. Die diskreten Punkte sind zu Darstellungszwecken zu einer kontinuierlichen Linie zusammengefasst.

Feedbackkomponente *E*

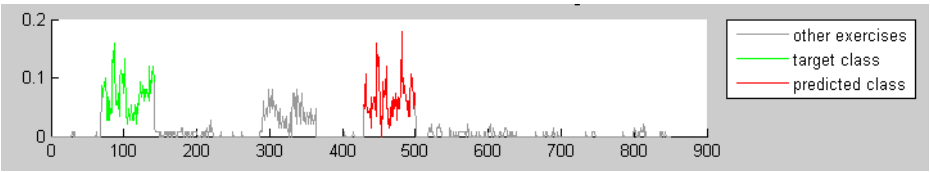
Im Fall des regionenbezogenen Feedbacks werden jeder der fünf lokalen Regionen zwei kontinuierliche Bewertungsmaße $\tilde{a_s}$ und $\tilde{v_s}$, mit $s \in \{1, ...,5\}$, zugeordnet. Diese bewerten die Übungsausführung des Patienten in Bezug auf die auszuführende Übung (Vorgabeübung). Um eine übersichtliche und intuitive Interpretation dieser mehrteiligen Information zu ermöglichen, ist eine kompakte Darstellung erforderlich. Zu diesem Zweck werden Diskretisierung und Farbschema von der globalen GUI-Komponente B übernommen und auf die schematische Abbildung eines Gesichts gemappt (siehe Tab. 5.4). Jede Region wird zudem mit der diskreten Übungsklasse beschriftet, die ihr vom lokalen Random-Forest zugeordnet wurde. Die Abbildung 5.26a zeigt die Ausgangsgrafik der lokalen GUI-Komponente E, die Abbildungen 5.26b und 5.26c visualisieren ein Beispiel des generierten Feedbacks. Eine ergebnisbezogene Analyse des Feedbacks erfolgt im Abschnitt 5.5.2.



(a)



(b)



(c)

Abbildung 5.25.: Beispielhafte Screenshots der Feedbackkomponente *D*. Diese zeigt einen Plot des Ähnlichkeitsvektors $\mathbf{a}_t$ (x-Achse: Index i der Trainingsobservation, y-Achse: paarweise Ähnlichkeit zwischen der Testobservation und der i-ten Trainingsobservation). Die diskreten Punkte sind zu Darstellungszwecken durch eine Linie verbunden. **(a)** Die Vorgabeklasse (VK) *Augen* wird vom Patient ausgeführt und vom System korrekt erkannt. Die (grün markierten) paarweisen Ähnlichkeiten zwischen der Testobservation und den Trainingsobservationen der VK sind deutlich größer als die übrigen paarweisen Ähnlichkeiten. **(b)** VK *Kuss* (sonst weitestgehend analog zu (a)). **(c)** Die Testobservation weist hohe Ähnlichkeiten zu den Trainingsobservationen der falsch-negativen VK *Kuss* (grün), der Klasse *UForm*, sowie der falsch-positiven Klasse *OForm* auf.

(a) (b) (c)

Abbildung 5.26.: **(a)** Ausgangsgrafik der Feedbackkomponente *E*. **(b)** - **(c)** Beispiel für die Visualisierung des lokalen Feedbacks für die Ausführung der Übung *Breit* durch einen Fazialisparesepatienten.

5.4.3. Technische Umsetzung und Laufzeiten

Der Prototyp wurde in der Interpretersprache Matlab (Version R2012a) umgesetzt. Er dient im Wesentlichen dazu, die Ergebnisse des entwickelten Feedbackverfahrens in anschaulicher Form zu visualisieren. um eine in ein reales Therapieszenario integrierbare Trainingsanwendung zu erhalten, wären weitere Anpassungen und Erweiterungen erforderlich. Zu diesen zählen unter anderem:

- die Portierung der Algorithmen in eine Sprache, die eine effiziente Programmierung erlaubt (z.B. C++).
- der Aufbau und die Gestaltung der grafischen Benutzeroberfläche in einer an die Zielgruppe angepassten Form (bspw. bezüglich der Usability).

Wie aus der Abbildung 5.1 ersichtlich wurde, lässt sich der Prozess der Feedbackerzeugung in vier Hauptkomponenten unterteilen. Diese umfassen die Vorverarbeitung der RGB- und Tiefendaten, die automatisierte Landmarkenlokalisierung, die Extraktion der Merkmalsdeskriptoren, sowie die Ableitung des Feedbacks aus den Merkmalsdeskriptoren (die eigentliche Erzeugung des Feedbacks). Jede Hauptkomponente umfasst verschiedene Teilkomponenten. Die gemessenen Laufzeiten der wichtigsten Teilkomponenten sind in der Tabelle 5.5 zusammengefasst (Rechner: Intel Core i7, 2 GHz, 8 GB). Somit werden in der Summe pro Bild 20 bis 30 Sekunden Laufzeit benötigt. Am rechenintensivsten ist die Initialisierung der Landmarkenlokalisierung, wobei die erforderliche Laufzeit abhängig von den gewählten Parametern ist und bei mehreren aufeinanderfolgenden Frames nur für den ersten Frame durchgeführt werden müsste. An zweiter Stelle folgt die Teilkomponente der Feedbackerzeugung mit einer Laufzeit von 6,9 Sekunden. Die Merkmalsextraktion hat demgegenüber geringe Laufzeitanforderungen.

Wie bereits erwähnt lag der Fokus dieser Arbeit auf der Entwicklung und Evaluierung des Feedbackverfahrens. Die Optimierung der Laufzeit war nicht primäre Zielstellung, insbesondere da sich für die Implementierung einer praxistauglichen Anwendung die Portierung in eine Compilersprache empfiehlt.

5.4.4. Zusammenfassung

In den vorhergehenden Abschnitten wurden zwei zentrale Themenpunkte der technischen Umsetzung behandelt. Der Schwerpunkt des ersten Teils lag im Wesentlichen auf den konzeptionellen Aspekten des Feedbackprototypen. Im Zuge dessen wurden die fünf Feedbackkomponenten der grafischen Benutzeroberfläche vorgestellt. Der zweite Teil umfasste eine Übersicht über die Details der technischen Umsetzung.

Teilkomponente	Laufzeit	Abschnitt
Vorverarbeitung		
Erzeugung 3D-Punktwolke (aus 2.5D-Bild)	25,7ms ± 1,6ms	3
Punktwolkenregistrierung (ICP)	4,7s ± 0,28s	4.2.1
Landmarkenlokalisierung		
Initialisierung [Asthana et al., 2014]	7s bis 16s	5.5.1
Optimierung [Asthana et al., 2014]	0,16s ± 0,02s	5.5.1
Merkmalsextraktion		
Distanz- & Winkelextraktion	3ms ± 0,26ms	4.3
Punktsignaturextraktion (8 Radien)	0,11s ± 0,08s	4.4
Krümmungmerkmalsextraktion	0,28s ± 0,03s	4.6
Feedbackerzeugung		
Feedbackerzeugung (global & lokal)	6,89s ± 0,16s	5.2.3, 5.2.4

Tabelle 5.5.: Übersicht über die Laufzeit ausgewählter Teilkomponenten. Die Berechnung erfolgte auf den Bilddaten einer zufällig gewählten Testperson (arithmetischer Mittelwert und Standardabweichung). Die Belegungen der merkmalsspezifischen Parameter entsprechen der des automatisiertern Testszenarios (siehe dazu Abschn. 5.5.1). Dementsprechend ist die HON-Merkmalsextraktion nicht eingebunden und die DW-Merkmale werden in reduzierter Form mit 14 Merkmalsvariablen extrahiert.

5.5. Experimentelle Evaluation prototypbezogener Aspekte

Die vorliegende Arbeit umfasst im Wesentlichen drei Experimentalteile, die auf verschiedene Aspekte der Mimiktrainer-Entwicklung bezogen sind. Der erste Teil in Kapitel 4 beschäftigte sich mit der ausführlichen Einzelevaluation der fünf gewählten Merkmalstypen. Im zweiten Teil wurde der kombinierte Einsatz der verschiedenen Merkmalstypen, die Random-Forest-Klassifikation, sowie die Ableitung der kontinuierlichen Bewertungsmaße untersucht (siehe Unterkap. 5.3).

Der folgende, dritte Experimentalteil ist stärker auf die Implementierung des Prototypen fokussiert und unterteilt sich in zwei Abschnitte. Der erste Abschnitt 5.5.1 untersucht den Einfluss der zu Grunde liegenden automatisierten Landmarkenlokalisierung auf die Erkennungsraten der zwölf Fazialisübungen und stellt die Ergebnisse denen der bisher verwendeten manuellen Landmarkenlokalisierung gegenüber. Die automatisier-

te Lokalisierung bildet eine wesentliche Voraussetzung für den Einsatz des entwickelten Verfahrens in einem realen Therapieszenario. Der folgende Abschnitt 5.5.2 wertet das resultierende diskrete Feedback der globalen und lokalen Feedbackerzeugung aus. Dabei werden sowohl Aufnahmen von gesunden Personen als auch von Fazialisparesepatienten evaluiert.

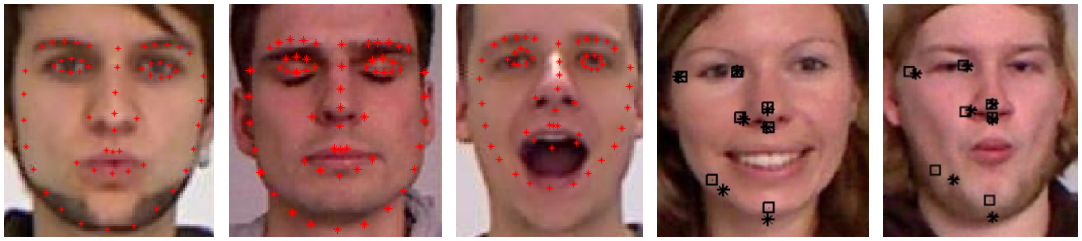
5.5.1. Automatisierung

In den vorhergehenden Experimentalabschnitten erfolgte die Bestimmung der Extraktionsareale auf Basis der manuell annotierten Landmarken der Ground-Truth. Bei einem Einsatz des Trainingssystems im Rahmen eines Echtzeit-Therapieszenarios ist ein manuelles Positionieren der Landmarken jedoch nicht praktikabel, weshalb im Zuge dieser Arbeit verschiedene Verfahren zur automatisierten Landmarkenlokalisierung evaluiert wurden.

Die von Cootes, Edwards und Taylor vorgestellten Active-Appearance-Models (AAMs) können unter anderem zur Lokalisierung von Landmarken eingesetzt werden [Cootes et al., 2001]. Dazu wurden in dieser Arbeit drei verschiedene Implementierungen von klassischen AAMs getestet und sowohl 2D-Intensitäts- als auch 2.5D-Tiefenbilder als mögliche Datenbasis evaluiert. Die erzielten Ergebnisse waren jedoch unzureichend, wie die Abbildungen 5.27a bis 5.27c beispielhaft zeigen. Da die Landmarken in dieser Arbeit dazu dienen, Areale für die Merkmalsextraktion innerhalb des Gesichts zu verorten und einzugrenzen, können sich Lokalisierungsfehler auch auf die Merkmalsextraktion und die nachfolgende Klassifikation der Fazialisübungen auswirken. Bei verschiedenen Experimenten zur Klassifikation von neun Fazialisübungen führte die AAM-basierte Landmarkenlokalisierung zu einer Verringerung der mittleren Erkennungsraten um 5 bis 30 Prozentpunkte gegenüber der Klassifikation mit manuell positionierten Landmarken. Nähere Details zu den Experimenten finden sich in [Lanz et al., 2013c] und [Dittmar et al., 2014].

Desweiteren wurde ein iteratives Lokalisierungsverfahren auf Basis von Krümmungsmerkmalen und distanzbezogenen Landmarkenrelationen entwickelt und evaluiert. Da sich die Krümmungseigenschaften der Gesichtsoberfläche in Folge der mimischen Bewegungen verändern, führte dies, insbesondere in den dynamischen Mund- und Wangenregionen, zu Abweichungen zwischen den manuell und automatisiert lokalisierten Landmarkenpositionen, wie die Abbildungen 5.27d und 5.27e verdeutlichen. Detailliertere Informationen und Ergebnisse zu diesem Lokalisierungsverfahren sind in [Lanz et al., 2013b] beschrieben.

Die präziseste Landmarkenlokalisierung erzielte das Verfahren von Asthana et al.,



(a) (b) (c) (d) (e)

Abbildung 5.27.: **(a)**–**(c)** Durch Anpassung eines Active-Appearance-Models (AAM) ermittelte Positionen von 58 Landmarken. Das Training der jeweiligen AAMs erfolgte auf den Intensitätsbildern von zehn Trainingspersonen, unter Ausschluss der Daten der dargestellten Testperson. (a) Korrektes Ergebnis. (b) Fehlende Anpassung an den Lidschluss, sowie ungenaue Lokalisierung des Kinnverlaufs. (c) Inkorrekte Lokalisierung des Mundes und der Gesichtssilhouette. **(d)**–**(e)** Lokalisierungsergebnisse des iterativen, krümmungs- und abstandsbasierten Verfahrens (Quadrate: Soll-Position, Sterne: Lokalisierungsergebnis). Die Abbildungen (d) und (e) sind aus [Lanz et al., 2013b] entnommen.

welches in [Asthana et al., 2014] näher beschrieben ist. Es basiert auf diskriminativen, deformierbaren Modellen. Die Lokalisierung erfolgt ausschließlich auf den Farbdaten. Eine in Matlab integrierbare Implementierung des Algorithmus wird, inklusive eines vortrainierten Modells, von den Entwicklern für wissenschaftliche Zwecke zur Verfügung gestellt[5] und in dieser Arbeit verwendet. Zur Initialisierung ist die Vorschaltung eines Gesichtsdetektors erforderlich, welcher ebenfalls auf der angeführten Internetseite bereitgestellt wird ([Zhu und Ramanan, 2012],[Asthana et al., 2013]). Auf dem 931 Aufnahmen umfassenden Datensatz dieser Arbeit erzielte dieses Entscheidungsbaum-basierte Gesichtsdetektionsverfahren Accuracy-, Precision- und Recall-Werte von 100 %. Nähere Details zu diesen Evaluationsmaßen sind unter anderem in [Sokolova und Lapalme, 2009] zusammengefasst. Im Gegensatz zur Gesichtsdetektion lässt sich die Genauigkeit der Landmarkenlokalisierung in Ermangelung einer Ground-Truth nicht sinnvoll evaluieren. So bezieht sich der Lokalisierungsansatz von Asthana et al., anders als die Ground-Truth dieser Arbeit, nur auf 49 statt 58 Landmarken. Desweiteren unterscheiden sich auch die zugeordneten Positionen innerhalb des Gesichts. So sind beim Ansatz von Asthana et al. im Vergleich mehr Landmarken im Bereich des Mundes definiert, jedoch keine entlang des Kinns. Da das vortrainierte Modell von Asthana et al. somit auf einer anderen Ground-Truth basiert,

[5] Chehra: Akshay Asthana, Stefanos Zafeiriou (`https://sites.google.com/site/chehrahome/`, letzter Zugriff: 28.08.2015).

ist ein direkter Vergleich der 58 manuell und 49 automatisiert gesetzten Landmarken, beispielsweise anhand euklidischer Abstände, zur Bewertung der Lokalisierungsgenauigkeit wenig aussagekräftig. Repräsentative Beispiele für die auf den Daten dieser Arbeit erzielten Lokalisierungen sind in den Abbildungen 5.28a bis 5.28f gezeigt. In Ermangelung korrespondierender, manuell gelabelter Landmarken, werden die automatisiert gefundenen sowohl für die Trainings- als auch die Testdaten zur Eingrenzung der Merkmalsextraktionsareale verwendet.

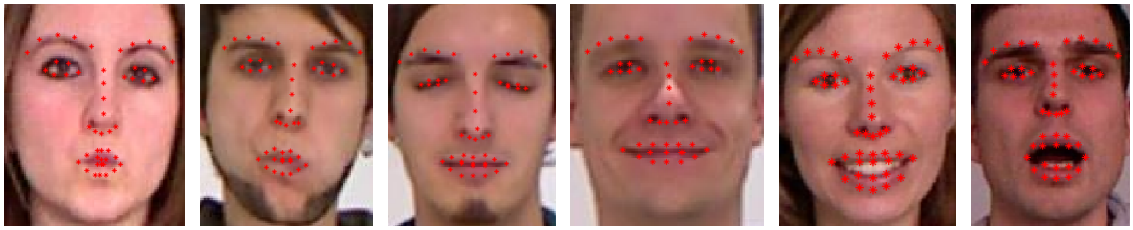
Weil die Unterteilung des Gesichts in zwölf Extraktionsregionen im Fall der patchbasierten Merkmalsextraktion anhand von rechteckigen Arealen erfolgt, können die unteren Extraktionsregionen Teile des Hintergrunds oder der Haare einschließen (vgl. Abb. 5.16a). Da eine Merkmalsextraktion aus diesen zu vermeiden ist, wurde bisher eine Segmentierung des Hintergrunds auf Basis der Kinn-Landmarken vorgenommen. Weil das von Asthana et al. bereitgestellte Lokalisierungsmodell jedoch keine Kinn-Landmarken zuordnet, wird der patchbasierten Merkmalsextraktion stattdessen eine wissens- und tiefendatenbasierte Vordergrund-Hintergrund-Segmentierung vorgeschaltet.

Aufgrund des reduzierten und veränderten Landmarkensatzes wird auch das Extraktionsverfahren der Distanz- und Winkelmerkmale angepasst und von 43 auf 14 Merkmalsvariablen reduziert. Die Vorgehensweise bei der Reduzierung orientiert sich zudem an den Ergebnissen der, im Abschnitt 4.3.2 dokumentierten, experimentellen Auswertung. Die zu extrahierenden Distanzen und Winkel sind in der Abbildung 5.29a visualisiert. Sie ergeben einen 14 Einträge umfassenden, reduzierten Merkmalsvektor $\mathbf{V}_r$:

$$\mathbf{V}_r = [\delta_6,\ \delta_7,\ \delta_9,\ \delta_{13},\ \delta_{14},\ \delta_{15},\ \delta_{10},\ \delta_{11},\ \theta_{11},\ \theta_{12},\ \theta_{15},\ \theta_{16},\ \theta_{22},\ \theta_{24}]\,. \tag{5.10}$$

Weiterführende Erläuterungen zu den einzelnen Einträgen des Merkmalsvektors finden sich in den Abbildungen 4.7a und 4.7b, sowie der Tabelle 4.3.

Um zu analysieren, ob und inwieweit eine zu Grunde liegende automatisierte Landmarkenlokalisierung das entwickelte Verfahren zur Feedbackerzeugung beeinträchtigt, wurde ein Teil der in Abschnitt 5.3.2 durchgeführten Experimente unter Einbindung der beschriebenen Änderungen wiederholt. Die Experimente beschränken sich in diesem Fall auf den Einsatz von Random-Forests, da allein diese für die Ableitung der paarweisen Ähnlichkeiten relevant sind. Mit Ausnahme der für die Distanz- und Winkelmerkmalsextraktion (DWM) beschriebenen Abwandlungen gelten weiterhin die in der Tabelle 5.1 aufgeführten Festlegungen. Die reduzierte Variante der DWM, noch

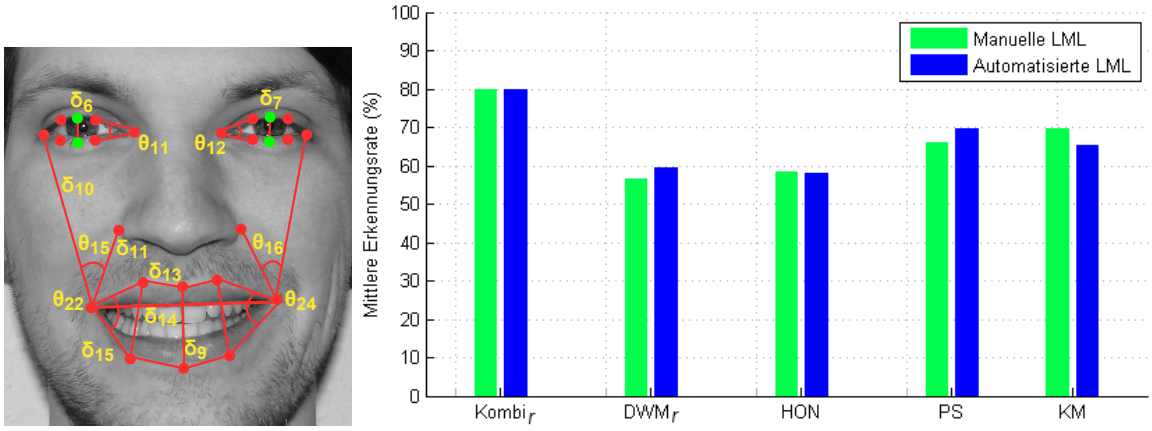


(a) (b) (c) (d) (e) (f)

Abbildung 5.28.: **(a)-(f)** Ergebnisse der automatisierten Lokalisierung von 49 Landmarken, basierend auf dem Verfahren von Asthana et al. [Asthana et al., 2014].

auf Basis der manuell gesetzten Landmarken, führte zu einer leichten Verschlechterung der mittleren Erkennungsraten (MER) im Vergleich zu den Ergebnissen aus Abschnitt 5.3.2 (Konkatenierung aller fünf Merkmalstypen: MER 79,80 % statt 81,03 %, Einzelevaluation DWM: MER 56,42 % statt 61,06 %). An dieser Stelle wären somit, als Ausblick für zukünftige Weiterentwicklungen, Anpassungen der zu extrahierenden Distanzen und Winkel sinnvoll. Dies sprengt jedoch den Rahmen dieser Arbeit, weshalb sich die folgenden Experimente auf den reduzierten Merkmalssatz beschränken.

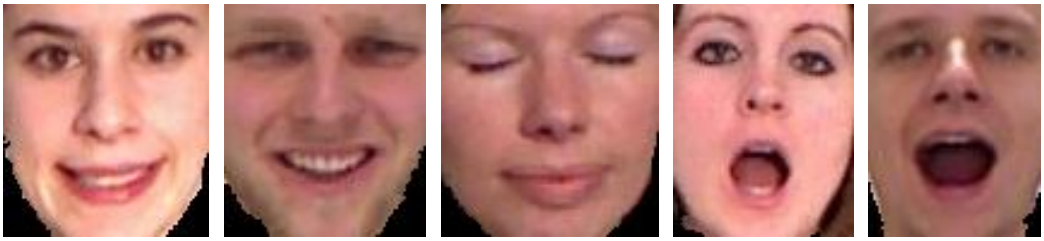
Die Ergebnisse der Random-Forest-Klassifikation von zwölf Fazialisübungen, mit zum einen manuell gesetzten und zum anderen automatisiert lokalisierten Landmarken, sind im Säulendiagramm 5.29b gegenübergestellt. Bei Konkatenierung der extrahierten Merkmalsvektoren aller fünf Merkmalstypen blieb die MER weitestgehend konstant (79,99 % statt 79,80 %), ebenso bei der Klassifikation mittels HON-Merkmalen, auf Basis welcher eine MER von 57,91 % erzielt wurde. Interessanterweise ergaben sich für die stärker landmarkenbasierten Merkmalstypen Verbesserungen der mittleren Erkennungsraten durch den Einsatz von automatisiert lokalisierten Landmarken. So verbesserte sich die MER für die alleinige Klassifikation mittels Punktsignaturen von 65,94 % auf 69,64 %. Die Klassifikation mittels DWM erzielte eine Verbesserung der MER von 56,42 % auf 59,41 %. Ein Grund hierfür könnte sein, dass das Setzen der Landmarken in automatisierter Form rein datenbasiert erfolgt und nicht, wie bei einem menschlichen Annotator, nach Augenmaß. Die aus dem alleinigen Einsatz von Krümmungsmerkmalen resultierende MER verschlechterte sich bei zu Grunde liegenden automatisiert lokalisierten Landmarken von 69,71 % auf 65,19 %. Einen potentiellen Grund hierfür könnte die, mangels Kinn-Landmarken, durchgeführte tiefendatenbasierte Segmentierung des Hintergrundes in den Extraktionsregionen der unteren Gesichtshälfte darstellen. Diese kann in einer ungenügenden Segmentierung resultieren, was zu einer unerwünschten Extraktion von Merkmalen aus Hintergrundarealen führt und die extrahierten Merkmalswerte verrauscht. Beispiele verschiedener



(a) (b)

Abbildung 5.29.: **(a)** Die auf Basis der automatisiert lokalisierten Landmarken zu extrahierenden Distanzen δ und Winkel θ. Die grün markierten Landmarken sind nicht Teil des Modells von Asthana et al. und werden als Mittelwert aus ihren linken und rechten Nachbarlandmarken bestimmt. **(b)** Säulendiagramm zur Gegenüberstellung der mittleren Erkennungsraten für zwölf Fazialisübungen mit zu Grunde liegender manueller bzw. automatisierter Landmarkenlokalisierung. Der Index r kennzeichnet den Einsatz des reduzierten DWM-Merkmalssatzes.

Ergebnisse der tiefendatenbasierten Vordergrund-Hintergrund-Segmentierung sind in den Abbildungen 5.30a bis 5.30e gezeigt. Für diese Erklärung spricht zudem, dass im Rahmen der vorhergehenden Experimente für die aus diesen Extraktionsregionen extrahierten Krümmungsmerkmalsvariablen die relativ gesehen höchste Mutual-Information geschätzt wurde (siehe dazu Abb. 4.37a). Bei den HON-Merkmalen erreichte die geschätzte Mutual-Information für die extrahierten Merkmalsvariablen der unterschiedlichen Extraktionsregionen einheitlichere Werte (vgl. Abb. 4.26). Möglicherweise stellen somit die aus den übrigen Regionen extrahierten Merkmalsdeskriptoren ausreichend Information bereit, um die mittlere Erkennungsrate weitesgehend konstant zu halten. So reduziert sich die MER bei zu Grunde liegenden automatisiert lokalisierten Landmarken für die HON-basierte Klassifikation nur leicht von 58,45 % auf 57,91 %. Eine eingehendere Überprüfung dieser Hypothese wäre durch entsprechende regionenspezifische Klassifikationstests möglich, sprengt jedoch den Rahmen dieser Arbeit.



(a) (b) (c) (d) (e)

Abbildung 5.30.: **(a)** - **(c)** Beispiele gelungener tiefendatenbasierter Vordergrund-Hintergrund-Segmentierung. Dem Hintergrund zugeordnete Pixel sind schwarz eingefärbt. Aus ihnen erfolgt keine Merkmalsextraktion. **(d)** - **(e)** Fehlgeschlagene Segmentierung, da Teile der Haare bzw. der Halsregion fälschlicherweise nicht dem Hintergrund zugeordnet wurden.

5.5.2. Evaluation des Feedbacks

In diesem Abschnitt wird das, vom Feedbacksystem erzeugte, diskrete Feedback abschließend evaluiert. Es umfasst sechs Feedbackstufen (FS 1 - FS 6), welche die Übereinstimmung der Patientenausführung mit den Modellausführungen der Trainingsdaten quantifizieren. Nähere Informationen zu den einzelnen Feedbackstufen sind unter anderem in der Tabelle 5.4 gegeben. Für jede Testobservation werden insgesamt sechs diskrete Bewertungsmaße bestimmt. Von diesen bezieht sich ein globales Bewertungsmaß auf das ganze Gesicht und fünf auf lokale Feedbackregionen (siehe dazu Abschn. 5.2.4).

Aus den in Unterkapitel 5.1 geschilderten Gründen, sind den Trainings- und Testaufnahmen dieser Arbeit keine Annotationen zur Qualität der dargestellten Übungsausführungen zugeordnet. Ein direkter Abgleich der durch das Feedbacksystem gewonnenen FS mit einer Ground-Truth ist daher nicht möglich. Infolgedessen wird eine indirekte Auswertung des Feedbacks durchgeführt. Die Auswertung unterteilt sich im Wesentlichen in zwei inhaltliche Blöcke, welche sich auf zwei verschiedene Datensätze stützen:

- Der *erste Auswertungsblock* beinhaltet eine quantitative Analyse und basiert auf dem bereits bekannten Datensatz. Dieser besteht aus 931 Aufnahmen von elf gesunden Personen und wurde in dieser Arbeit im Rahmen der Entwicklung sowohl als Trainings-, Validierungs- und Testdatensatz verwendet (wobei jedoch immer auf eine personenbezogene Kreuzvalidierung geachtet wurde).

- Ergänzend dazu werden im *zweiten Auswertungsblock* einzelne Beispielausführungen evaluiert und gegenübergestellt. Der Schwerpunkt liegt dabei auf den

Aufnahmen von Fazialisparesepatienten. Diese Aufnahmen sind Teil eines kleineren je 117 Tiefen- und RGB-Bilder umfassenden Datensatzes, welcher als reiner Testdatensatz Anwendung findet. Dies bedeutet, dass er, ebenso wie die aufgenommenen fünf Fazialisparesepatienten, zu keinem Zeitpunkt und in keiner Form in den Entwicklungsprozess des Feedbackverfahrens einbezogen wurde. Aus diesem Grund ist der Datensatz dazu geeignet, die Generalisierbarkeit des entwickelten Verfahrens zu evaluieren.

Ausführliche Informationen zu beiden Datensätzen finden sich im Abschnitt 2.3.2.

Quantitative Auswertung

Der erste Auswertungsblock entspricht im Prinzip einem realen Trainingsszenario. Bei diesem wird eine Zielübung vom Feedbacksystem vorgegeben, die tatsächlich vom Patienten ausgeführte Mimikbewegung ist jedoch nicht bekannt.

Im ersten Schritt der Auswertung wird dazu eine Zielübung k festgelegt. Für die folgenden quantitativen Analysen sind dies beispielhaft die Übungen *Kuss* und *Breit*. Anschließend werden, zusätzlich zu den Testobservationen der Zielübung k, auch Testobservationen anderer Übungsklassen an das Feedbacksystem übergeben[6]. Das System bewertet alle Testobservationen hinsichtlich der Zielübung k, unabhängig von ihrer tatsächlichen Klassenzugehörigkeit, welche dem System auch nicht bekannt ist. Im Folgenden werden die zugewiesenen Feedbackstufen jedoch klassenweise, also geordnet nach den Klassenzugehörigkeiten der Testobservationen, analysiert.

Die global zugewiesenen Feedbackstufen sind in der Abbildung 5.31 ersichtlich. Für die Analyse des lokalen Feedbacks wurden beispielhaft die Feedbackregionen 1 und 3 herausgegriffen (siehe Abb. 5.16c). Die quantitative Auswertung der zugewiesenen lokalen Feedbackstufen ist in den Schaubildern 5.32a bis 5.32d gezeigt. In allen Schaubildern orientiert sich die Zuordnung der Farben zu den einzelnen Feedbackstufen an dem bereits vorgestellten Farbschema, wobei grün eine hohe und rot eine niedrige Übereinstimmung mit der angestrebten Übungsausführung kennzeichnet (vgl. Tab. 5.4).

In beiden Diagrammen der Abbildung 5.31 wird ersichtlich, dass die Testobservationen, die tatsächlich der jeweiligen Vorgabeübung angehören, insgesamt auch am besten bewertet werden. Im Detail erhalten circa 90 Prozent dieser Testobservationen eine Bewertung, die mindestens der Feedbackstufe 3 entspricht. Übergibt man dem System, bei angenommener Zielübung *Kuss*, Testobservationen der Übungen *UForm* und

[6]Die Trainingsdaten umfassen zehn Personen, die Testdaten die verbliebene elfte Person. Über eine 11-fache personenbezogene Kreuzvalidierung fungiert jede Person des Datensatzes einmal als Testperson. Die Ergebnisse aller elf Testpersonen werden anschließend zusammengefasst.

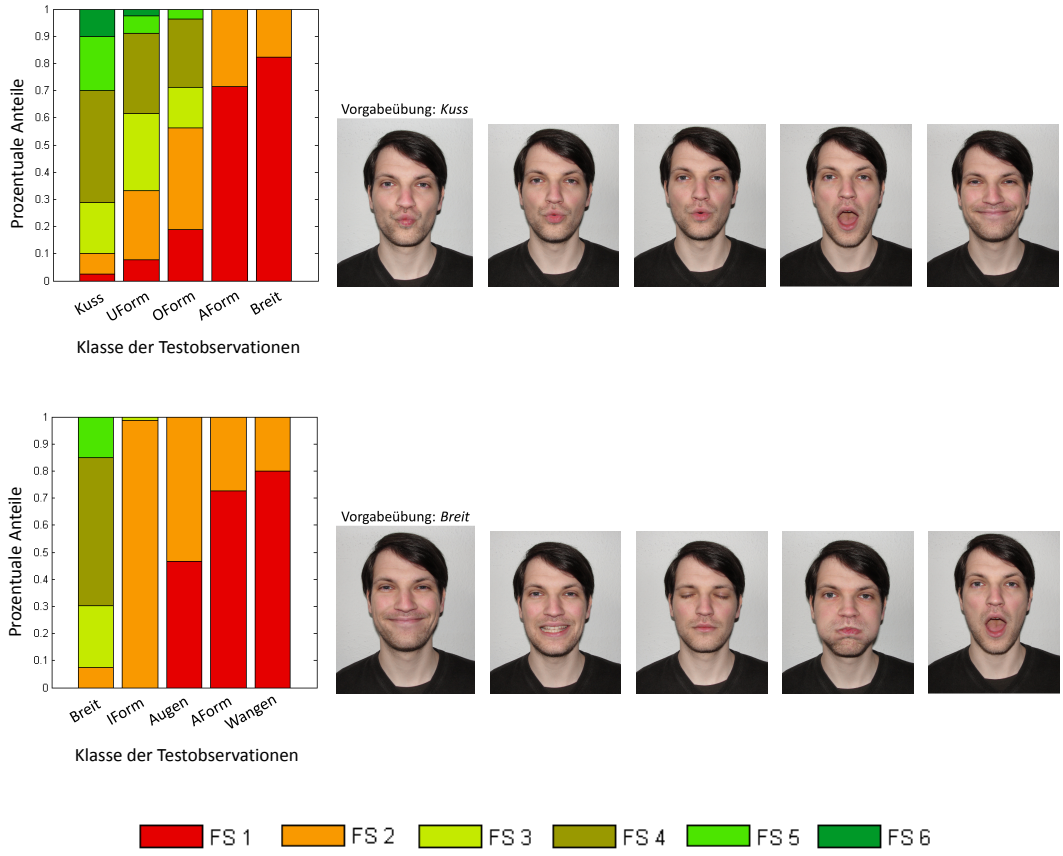


Abbildung 5.31.: Prozentuale Anteile der global zugewiesenen Feedbackstufen für Testobservationen von fünf verschiedenen Übungsklassen. **Zeile 1:** Vorgabeübung *Kuss*. **Zeile 2:** Vorgabeübung *Breit*. **Zeile 3:** Farblegende der sechs diskreten Feedbackstufen (FS). Nähere Details zu den einzelnen FS finden sich in der Tabelle 5.4.

OForm trifft dies auf 65 bzw. 40 Prozent der Testobservationen zu (siehe Abb. 5.31, Zeile 1). Alle drei Übungen weisen eine hohe ausführungsbezogene Ähnlichkeit durch die Aktivierung des Mundringmuskels auf, wobei jedoch subjektspezifische Unterschiede existieren können (vgl. Abb. 5.13). Ein weit geöffneter Mund (*AForm*) oder die laterale Verschiebung der Mundwinkel (*Breit*) resultiert, bei gegebener Zielübung *Kuss*, für 100 Prozent der evaluierten Testobservationen in einer niedrigen Bewertung (FS 1 oder 2).

Die Zielübung *Breit* unterscheidet sich in ihrer Ausführung deutlich von den gewählten Vergleichsübungen (*IForm*, *Augen*, *Wangen* und *AForm*). Dies spiegelt sich auch in den zugewiesenen Feedbackstufen wider (siehe Abb. 5.31, Zeile 2). Die relativ gesehen besten Bewertungen erhalten die Testobservationen der Übung *IForm* (FS 2, teilw. FS 3). Sowohl die Übung *Breit* als auch die Übung *IForm* zeichnen sich durch

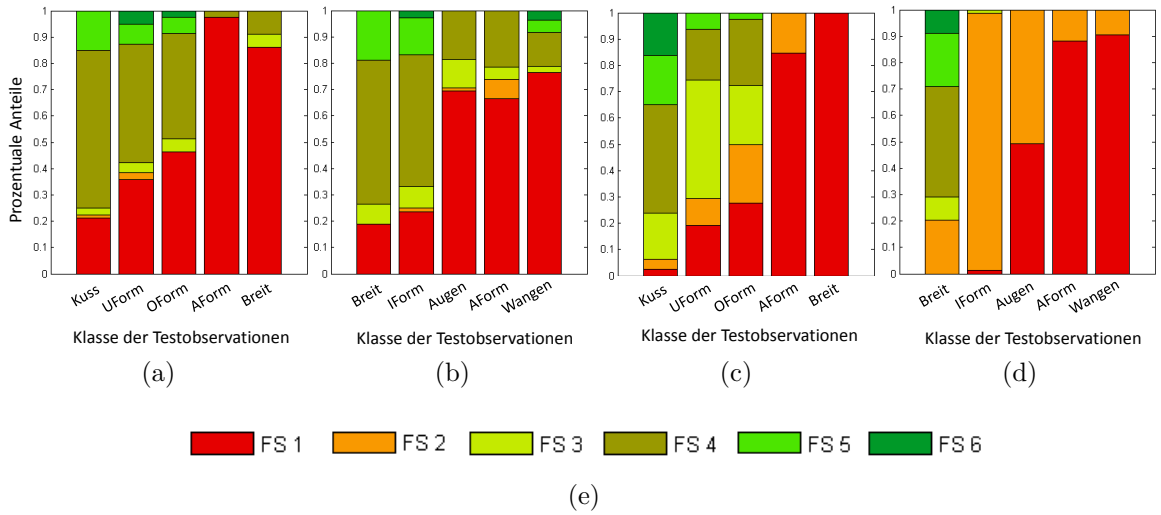


Abbildung 5.32.: Prozentuale Anteile der zugewiesenen regionenbezogenen Feedbackstufen am Beispiel von zwei verschiedenen Regionen. Die Vorgehensweise der Auswertung entspricht im Wesentlichen der globalen Vorgehensweise (vgl. Abb. 5.31). Die Extraktion der Merkmalsdeskriptoren, welche der Feedbackableitung zu Grunde liegen, beschränkt sich jedoch auf **(a)**-**(b)** die Augenregion 1 bzw. **(c)**-**(d)** die Wangenregion 3. Während das für die Wangenregion ermittelte Feedback grundsätzlich dem Globalen ähnelt, weist das auf die Augenregion bezogene Feedback ein höheres Rauschen auf. **(e)** Farblegende der sechs diskreten Feedbackstufen.

eine laterale Verschiebung der Mundwinkel, unter anderem durch die Aktivierung des Musculus risorius, aus[7].

Zur Evaluation der lokal zugewiesenen Feedbackstufen werden im Folgenden beispielhaft die Ergebnisse für die Augenregion 1 (R1) und die Wangenregion 3 (R3) vorgestellt und verglichen (Abb. 5.32a bis 5.32d). Um einen Vergleich mit den Ergebnissen der globalen Feedbackevaluation zu ermöglichen, werden die Fazialisübungen *Kuss* und *Breit* erneut als Vorgabeübungen festgelegt. Die prozentualen Verteilungen des globalen und des R3-bezogenen Feedbacks stimmen im Wesentlichen überein und sind, wie in den vorhergehenden Absätzen deutlich wurde, durch anatomische Un-/Ähnlichkeiten in der Übungsdurchführung erklärbar. Das auf die Augenregion bezogene Feedback unterscheidet sich von diesen und weist zudem einzelne Unstimmigkeiten auf. So ist in der Abbildung 5.32a ersichtlich, dass 5 bzw. 2 Prozent der *UForm*- und *OForm*-Testobservationen mit der bestmöglichen Feedbackstufe 6 be-

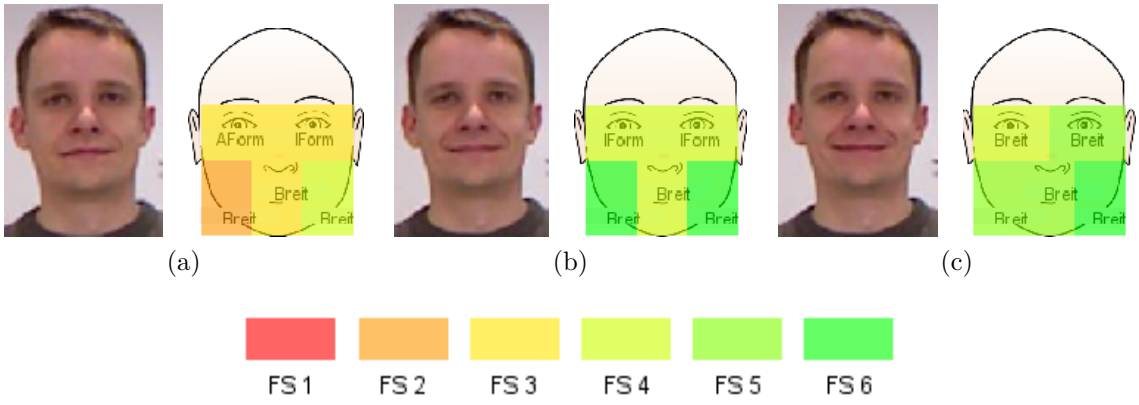
[7]Anschauliche Demonstrationen zur Funktion der einzelnen Muskeln sind verfügbar unter: `http://flexikon.doccheck.com/de/Mimische_Muskulatur`; (letzter Zugriff: 01.03.2016)); DocCheck Medical Services GmbH, Dr. Frank Antwerpes.

wertet werden. Testobservationen der Vorgabeübung *Kuss*, deren Trainingsdaten den Orientierungsrahmen für die Feedbackableitung bilden, werden demgegenüber maximal mit der Feedbackstufe 5 bewertet. Gleiches zeigt sich in der Abbildung 5.32b für die Testobservationen der Übungen *Breit* (Vorgabeübung) und *IForm*. Vor dem Hintergrund, dass ein Lidschluss bei der Fazialisübung *Breit* nicht erwünscht ist, ist zudem die Zuweisung der Feedbackstufen 3 und 4 zu insgesamt 30 Prozent der *Augen*-Testobservationen als fehlerhaft anzusehen. In Übereinstimmung mit den experimentellen Ergebnissen in Abschnitt 5.3.3 bestätigt dies die Notwendigkeit für eine weitere Optimierung der lokalen Merkmalsextraktion, um eine bessere Repräsentation der trainings- und testdateneigenen Strukturen zu ermöglichen.

Qualitative Gegenüberstellung von Einzelergebnissen

Die quantitative Feedbackevaluation des vorhergehenden Abschnitts stützte sich auf die Klassenzugehörigkeit der Testobservationen. Ergänzend dazu erfolgt in diesem Abschnitt eine qualitative Auswertung des resultierenden Feedbacks auf Basis von Beispielaufnahmen der Vorgabeklasse.

In den Abbildungen 5.31 (Zeile 2), 5.32b und 5.32d wurde ersichtlich, dass den Testobservationen der Vorgabeklasse *Breit* global und lokal alle möglichen Feedbackstufen zugewiesen wurden. Zur Veranschaulichung sind in den Abbildungen 5.33a bis 5.33c drei Beispielaufnahmen einer gesunden Testperson, einschließlich des zugewiesenen Feedbacks, gezeigt. Es ist erkennbar, dass den Ausführungen, die eine höhere Intensität in der Übungsdurchführung zeigen, höhere globale und lokale Feedbackstufen (FS) zugewiesen wurden. Zudem wird deutlich, das lokale Veränderungen in ihren Auswirkungen nicht isoliert zu betrachten sind. Eine zunehmende laterale Verschiebung der Mundwinkel wirkt sich beispielsweise auf die Wangenwölbung in Jochbeinhöhe (Feedbackregion 1 und 2) und die Spannung der Lippen in der Mundregion (Feedbackregion 4) aus. Die einer lokalen Regionen zugewiesene Feedbackstufe (Farbe) bezieht sich immer auf die vom Patienten auszuführende Vorgabeklasse, welche in der gezeigten Abbildung der Übung *Breit* entspricht. Der zugeordnete Übungsname kennzeichnet das lokale Klassifikationsergebnis. Die Farbe einer Region ist, im Fall einer lokalen Falschklassifikation, somit nicht mit der beschrifteten Übungsklasse verknüpft. Bei der gezeigten Testperson handelt es sich um eine gesunde Person, die Aufnahmen sind zudem Teil des Entwicklungsdatensatzes. Um zusätzlich die Generalisierbarkeit des entwickelten Verfahrens analysieren zu können, basierend die weiteren Auswertungen auf den Aufnahmen von Fazialisparesepatienten. Bei diesen handelt es sich um reine Testdaten. Dies bedeutet, dass sie zu keinem Zeitpunkt als Trainings-,



(a) (b) (c)

(d)

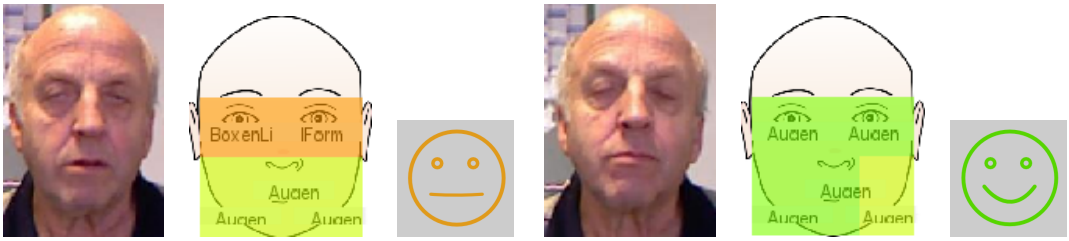
Abbildung 5.33.: Ausführung der Übung *Breit* durch eine gesunde Testperson. Die global zugeordneten Bewertungsstufen sind: **(a)** FS 3 **(b)** FS 4 **(c)** FS 4. Die zunehmende laterale Verschiebung der Mundwinkel wirkt nicht isoliert in den Feedbackregionen 3 und 4 (FR), sondern geht mit Veränderungen in allen Feedbackregionen einher (Wangenwölbung in FR 1 und 2, Spannung der Lippen (FR 4)). Die Farbe der lokalen Regionen bezieht sich immer auf die Vorgabeklasse (hier: *Breit*), die zugeordneten Übungsnamen beschreiben das lokale Klassifikationsergebnis. **(d)** Farblegende für die Einfärbung der lokalen Regionen.

Validierungs- oder Testdatensatz in die Entwicklung des Verfahrens einbezogen wurden.

Die Abbildungen 5.34a und 5.34b umfassen zwei unterschiedlich Ausführungen der Übung *Augen* durch einen Fazialisparesepatienten. In der Gegenüberstellung der Aufnahmen ist ein sinnvolles Feedbackresultat erkennbar. Isoliert betrachtet ist anzumerken, dass die FS 4 in Anbetracht des leicht geöffneten Mundes für die Wangen- und Mundregion eine möglicherweise zu hohe Bewertung darstellt (Farblegende siehe Abb. 5.33d). Für einen Vergleich mit der Übungsausführung einer gesunden Person wird für diese und die folgenden Übungen auf die Abbildung 2.6 verwiesen.

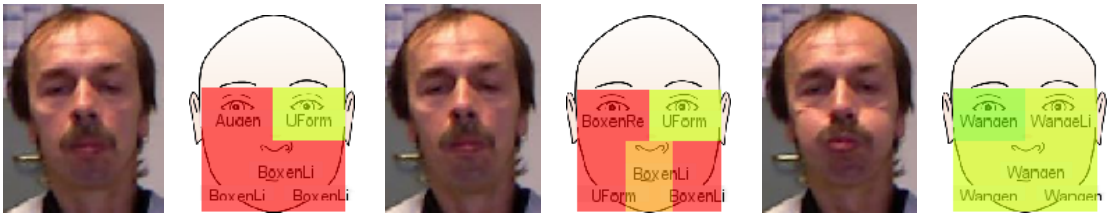
Die Abbildungen 5.35a bis 5.35c enthalten Beispielaufnahmen für die Vorgabeübung *Wangen* zu unterschiedlichen Zeitpunkten. Die ersten beiden Ausführungen entsprechen Zwischenschritten (global FS 1), die letzte Abbildung zeigt die finale Version (FS 4). Beim Vergleich der Aufnahmen zeigt sich ein überwiegend konsistenter Verlauf des lokalen und globalen Feedbacks. Eine Ausnahme bildet die Bewertung der lokalen Augenregion 2 in der ersten und zweiten Abbildung mit der Feedbackstufe 4.

Die Abbildungen 5.36a bis 5.36c zeigen verschiedene Beispiele für die Ausführung der Übung *BoxenRe* durch einen Fazialisparesepatienten. Der nicht erwünschte Augen-



(a) (b)

Abbildung 5.34.: Ausführung der Übung *Augen* durch einen Fazialisparesepatienten. **(a)** Unzureichende Übungsausführung mit unvollständigem Lidschluss und leicht geöffnetem Mund (globale FS 2). **(b)** Mit Ausnahme eines leicht herabhängenden Mundwinkels eine weitestgehend korrekte Übungsausführung (globale FS 5).



(a) (b) (c)

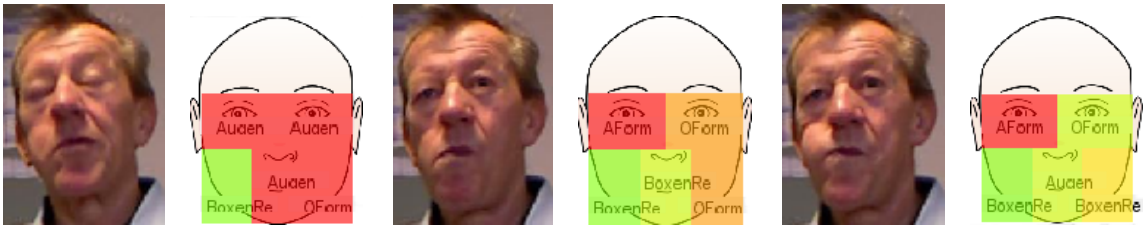
Abbildung 5.35.: Ausführung der Vorgabeübung *Wangen* durch einen Fazialisparesepatienten mit zugeordnetem lokalem Feedback und Klassifikationsergebnis (Farblegende siehe Abb. 5.33d). Die global zugeordneten Feedbackstufen sind: **(a)** FS 1 **(b)** FS 1 **(c)** FS 4.

schluss wird korrekt erkannt, ebenso der mangelnde Einbezug der Mundmuskulatur. Das lokale Klassifikationsergebnis *Augen* für die Mundregion 3 erscheint im ersten Augenblick ungewöhnlich (Abb. 5.36a und 5.36c). Ein Vergleich der Mundregionen in den Abbildungen 2.6c und 2.6e verdeutlicht jedoch den Unterschied zwischen den leicht gespitzen Lippen der Fazialisübung *BoxenLi* und dem neutralen Mund der Übung *Augen*.

Weitere Feedbackbeispiele sind in den Abbildungen 5.37 und 5.38 gezeigt.

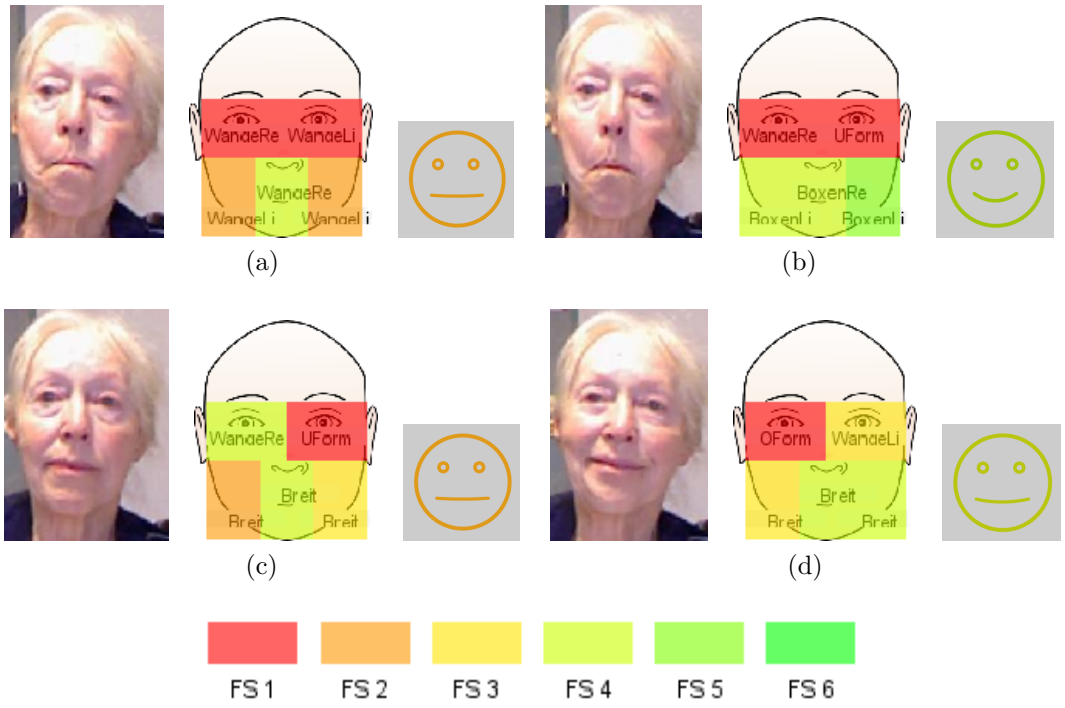
5.6. Zusammenfassung und Ausblick

Der Fokus dieses abschließenden inhaltlichen Kapitels lag auf der Feedbackableitung und der Umsetzung eines Prototypen. Die Gliederung des Kapitels lässt sich im Wesentlichen in drei thematische Blöcke unterteilen. Im ersten, theoriebezogenen Block wurden potentielle Vorgehensweisen und Verfahren zur Ableitung von Feedback



(a) (b) (c)

Abbildung 5.36.: Beispiele für die Ausführung der Übung *BoxenRe* durch einen Fazialisparesepatienten. Die global zugewiesenen Feedbackstufen sind: **(a)** FS 2 **(b)** FS 2 **(c)** FS 3.



(a) (b) (c) (d) (e)

Abbildung 5.37.: Übungsausführungsbeispiele einer Fazialisparesepatientin mit zugeordnetem lokalem und globalem Feedback. **(a)-(b)** Vorgabeübung *BoxenLi*: Im ersten Beispiel ist die Zunge zu nah an der Nase und zu wenig lateral positioniert. Neben der lokalen auch globale Falschklassifikation (*WangeLi*). Die rechte Abbildung zeigt eine bessere Ausführung. Die Übungsklasse *BoxenLi* wurde auch global korrekt erkannt. **(c)-(d)** In beiden Fällen unzureichende laterale Verschiebung der Mundwinkel bei der Durchführung der Übung *Breit* (FS 2 und 3). **(e)** Legende zur Einfärbung der Regionen.

aus extrahierten Merkmalsdeskriptoren diskutiert. Anschließend wurde im zweiten Block die entwickelte Methode vorgestellt und experimentell evaluiert (Unterkap. 5.2 und 5.3). Der Schwerpunkt der Evaluation lag dabei auf dem Verfahren und seinen Teilkomponenten. Demgegenüber war der Fokus des dritten und abschließenden Blocks stärker ergebnisbezogen. Im Unterkapitel 5.4 fanden sich Eckdaten zum implementierten Prototypen, welcher zur Veranschaulichung des ermittelten Feedbacks dient. Die Umsetzung einer Feedbacksoftware setzt zudem eine Automatisierung der Landmarkenlokalisierung voraus, die, ebenso wie die Auswertung des resultierenden globalen und regionenbezogenen Feedbacks, Gegenstand der experimentellen Evaluation in Unterkapitel 5.5 war. Die quantitative und qualitative Evaluation des resultierenden diskreten Feedbacks bestätigt die grundsätzliche Eignung des vorgestellten Ansatzes für ein automatisiertes Feedbackszenario. Für den Ausblick ergeben sich drei konkrete weiterführende Schritte. Diese umfassen:

- eine weitere Verbesserung der lokalen Merkmalsextraktion durch geeignete Merkmalsextraktionsverfahren (siehe dazu Kap. 4 und 6).
- eine systematische Evaluation des erzeugten Feedbacks durch Experten (z.B. Logopäden, Sprechwissenschaftler) und Patienten.
- die Ergänzung des Verfahrens um ein interaktives Reinforcement-Learning, basierend auf (optionalem) Feedback des Logopäden. Das in dieser Arbeit entwickelte Verfahren funktioniert merkmalsbasiert, sodass eine vorgeschaltete zeitintensive und komplexe manuelle Annotation der Trainingsdaten hinsichtlich der Übungsqualität nicht erforderlich ist (Hinweise zu Problemstellungen der manuellen Annotation siehe Unterkap. 5.1). Um diesen Vorteil beizubehalten und dennoch Expertenwissen einzubeziehen, wäre die Integration einer einfach gehaltenen und optional zu bedienenden Feedbackschnittstelle für den Logopäden sinnvoll.

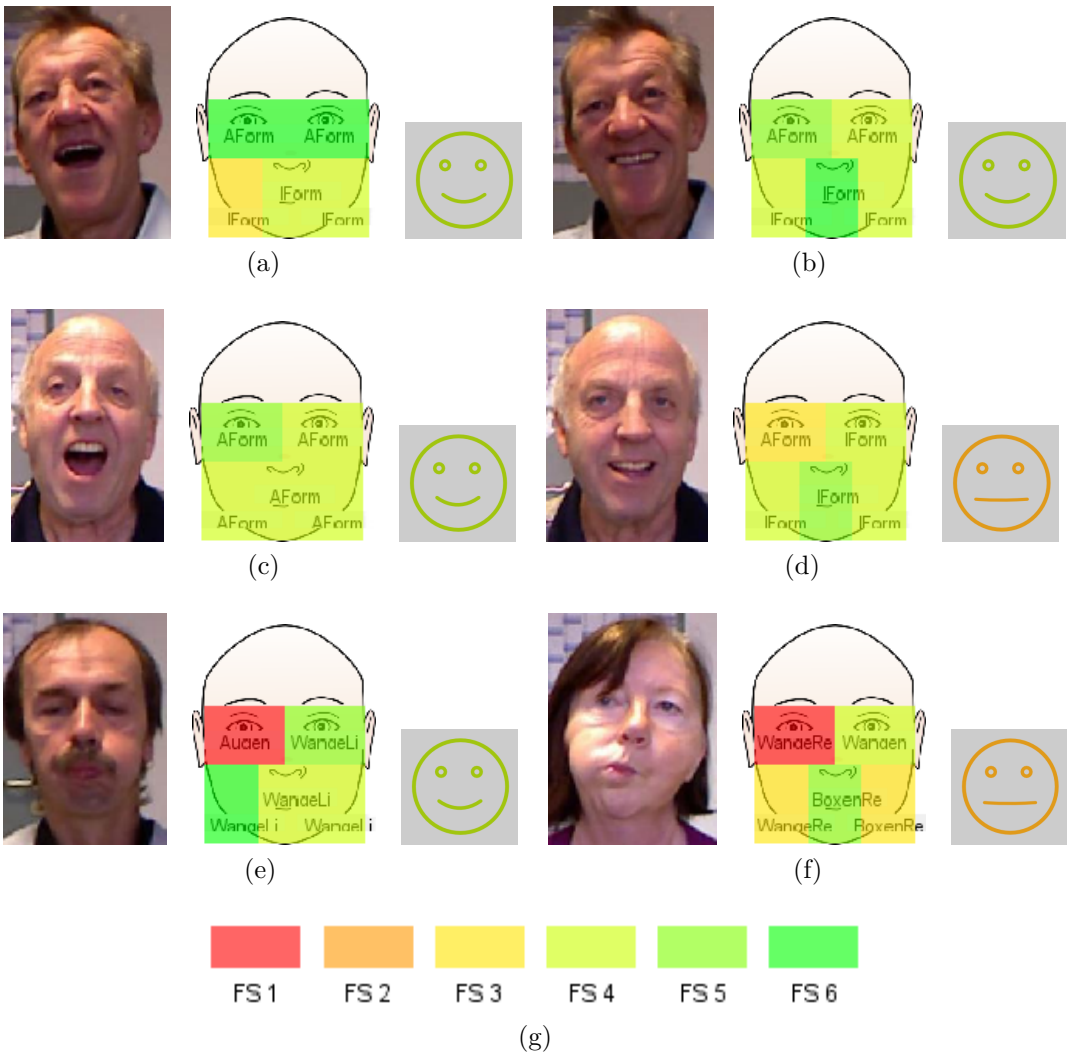


Abbildung 5.38.: Übungsausführungsbeispiele für fünf Fazialisparespatienten mit zugeordnetem globalen und lokalen Feedback. Die Vorgabeübungen sind **(a)** *AForm*, **(b)** *IForm*, **(c)** *AForm*, **(d)** *IForm*, **(e)** *WangeLi* und **(f)** *BoxenRe*. **(f)** Farblegende für das lokale Feedback.

6. Abschließende Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde ein Prototyp für die Feedbackkomponente eines therapiebegleitenden Mimiktrainers konzipiert und umgesetzt. Der Aufbau dieser schriftlichen Ausarbeitung orientierte sich weitestgehend an der zugrunde liegenden technischen Gesamtarchitektur, welche in der Abbildung 6.1 gezeigt ist. Die Architekturkomponenten lassen sich im Wesentlichen in vier thematische Schwerpunkte einordnen, die in einzelnen Kapiteln erörtert und evaluiert wurden. In den folgenden vier Abschnitten werden die wesentlichen Inhalte und Erkenntnisse dieser Kapitel zusammengefasst. Im Sinne eines Ausblicks finden sich ergänzend Hinweise auf Anknüpfungspunkte für zukünftige weiterführende oder themennahe Arbeiten.

Anwendungsszenario und Gesamtarchitektur

Der Fokus von Kapitel 2 lag auf der Vorstellung des Anwendungsszenarios, sowie einer einleitenden Übersicht über die einzelnen Komponenten der technischen Gesamtarchitektur. Um den konzeptionellen Beitrag dieser Arbeit aufzuzeigen, wurde zu Beginn eine Übersicht über den Stand der Wissenschaft zur rehabilitationsbezogenen, computergestützten Mimikanalyse gegeben. Die Übersicht stützt sich auf insgesamt 27 Veröffentlichungen aus den Jahren 2000 bis 2016. In einer deutlichen Mehrheit der vorgestellten Arbeiten konzentrieren sich die Autoren auf die Entwicklung von Verfahren zur Diagnose von Mimikdysfunktionen. Lediglich fünf Publikationen sind, wie die auch die vorliegende Arbeit, auf die Vorstellung und Entwicklung von therapiebegleitenden Trainings- und Feedbacksystemen ausgerichtet. Eine detaillierte Übersicht über die konzeptionellen und anwendungsszenariobezogenen Beiträge dieser Arbeit wurde im Unterkapitel 1.2 gegeben.

Da im Rahmen der Literaturrecherche keine geeignete Datenbasis für die Entwicklung und Evaluierung des Feedbackverfahrens ausfindig gemacht werden konnte, wurden, unterstützt durch Birant Sibel Olgay im Rahmen ihrer Masterarbeit [Olgay,

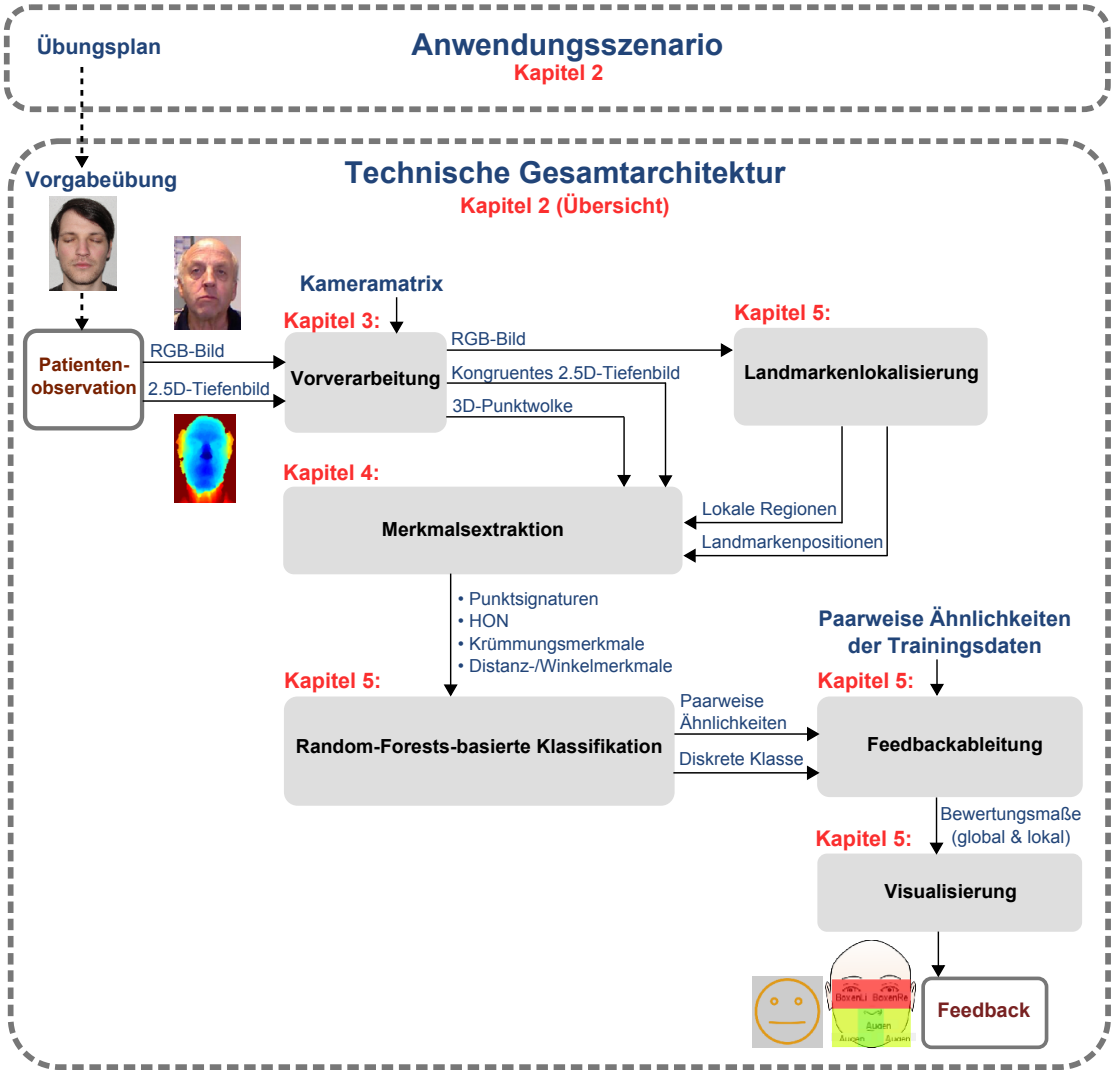


Abbildung 6.1.: Überblick über die technische Gesamtarchitektur des in dieser Arbeit entwickelten Verfahrens zur Feebackerzeugung.

2012][1], zwei Datensätze gesammelt. Diese umfassen RGB- und 2.5D-Tiefenaufnahmen von gesunden Personen bzw. Fazialisparese-Patienten bei der Durchführung von therapeutischen Fazialisübungen. Details zu beiden Datensätzen und den gewählten Fazialisübungen sind im Unterkapitel 2.3 gegeben.

Datenaufnahme und -vorverarbeitung

Die wesentlichen Schritte der Datenaufnahme und -vorverarbeitung wurden in Kapitel 3 dokumentiert und beruhen, mit Ausnahme der in Unterkapitel 3.4 beschriebenen Vorgehensweise, auf bekannten und grundlegenden Methoden. Für einmalig auszufüh-

[1]Diese Masterarbeit wurde von der Autorin im Rahmen dieser Dissertation betreut.

rende Arbeitsschritte, wie z.B. die Schätzung der Kameraparameter (Kalibrierung), wurde daher auf von Dritten zur Verfügung gestellte Funktionsbibliotheken zurückgegriffen. Nähere Details zu letzteren sind in den jeweiligen Abschnitten gegeben. Die für den Ablauf der Anwendung relevanten, wiederkehrenden Operationen zur Korrespondenzermittlung und Punktwolkengenerierung wurden hingegen im Rahmen dieser Arbeit nachimplementiert. Auf diese Weise funktioniert die Vorverarbeitung auch ohne Einbindung umfangreicher, teilweise Library- und Betriebssystem-abhängiger, Funktionsbibliotheken (z.B. Point Cloud Library, Kinect SDK).

Merkmalsextraktion

Die extrahierten Merkmalsdeskriptoren sind ein wesentlicher Bestandteil der Feedbackerzeugung. Aus diesem Grund stellt die in Kapitel 4 enthaltene Auswahl und Evaluation der Deskriptoren einen zentralen Themenpunkt dieser Arbeit dar.

Im Unterkapitel 2.1 wurden bereits verschiedene Veröffentlichungen zur computergestützten Rehabilitation von Mimikdysfunktionen vorgestellt. Die beschriebenen Vorgehensweisen und Merkmalsextraktionsverfahren basierten jedoch überwiegend auf Annahmen, die für das Szenario dieser Arbeit nicht zutreffend oder sinnvoll sind (z.B. achsensymmetrische Fazialisübungen, auf einzelne Übungen zugeschnittene Merkmalsextraktion). Die letztendlich gewählten Merkmalsextraktionsverfahren haben ihren Ursprung daher nicht im konkreten Anwendungsszenario. Vielmehr handelt es sich bei ihnen um allgemeine Ansätze aus dem Themenfeld der automatisierten Gesichtsanalyse. Die Ergebnisse der dazugehörigen Literaturrecherche sind im Unterkapitel 4.1 zusammengefasst (vgl. auch Tab. B.3). Aufbauend auf den Ergebnissen der Recherche wurden fünf verschiedene Merkmalsextraktionsverfahren ausgewählt, bestehend aus Punktsignaturen, Histogrammen orientierter Normalenvektoren, sowie Distanz-, Winkel- und Krümmungsmerkmalen. Sie wurden in den Unterkapiteln 4.3 bis 4.6 gesondert vorgestellt und einer eingehenden experimentellen Evaluation unterzogen. Ein Überblick über die Einzelergebnisse erfolgte im Abschnitt 4.7, ergänzt um einen Leitfaden zur Merkmalsauswahl. Aus den Ergebnissen der Evaluationen ergeben sich zudem konkrete Anknüpfungspunkte für weiterführende Arbeiten:

- Im Originalverfahren von [Rabiu et al., 2012] werden die aus der linken und rechten Gesichtshälfte extrahierten Distanzen teilweise gemittelt (siehe Unterkap. 4.3). An dieser Stelle wäre zu untersuchen, ob und inwieweit diese Vorgehensweise möglicherweise relevante Informationen über die Achsensymmetrie des Gesichts auslöscht. Für diese Annahme spricht, dass die extrahierten Winkelmerkmalsdeskriptoren nicht gemittelt werden und trotz geringerer Anzahl an

Merkmalsvariablen (16 vs. 27) im Rahmen der Klassifikation zu höheren mittleren Erkennungsraten für die Fazialisübungen führen.

- Ausdehnung der in Unterkapitel 4.4 evaluierten Punktsignaturmerkmalsextraktion auf weitere Gesichtsareale durch Verschieben des Punktsignaturzentrums von der Nasenspitze in andere Landmarken. Erste Analysen dazu mit Signaturzentren auf dem Nasenrücken bzw. neben den Nasenflügeln finden sich in der Masterarbeit von Birant Olgay [Olgay, 2012] (basierend auf manuell gesetzten Landmarken und einer Vorgängerversion des in dieser Arbeit erweiterten Punktsignaturextraktionsverfahrens). Durch das robuste Lokalisierungsverfahren von Asthana et al. [Asthana et al., 2014] ist jedoch eine Erweiterung selbst auf Landmarken in sehr dynamischen Gesichtsarealen möglich (siehe Abschn. 5.5.1). Erste Analysen, die jedoch nicht in dieser Arbeit dokumentiert sind, zeigten vielversprechende Ergebnisse für den Einsatz der Mundwinkel als Signaturzentrum.

Aufgrund der Fülle an existierenden Verfahren zur automatisierten Gesichtsanalyse, beschränkt sich die Literaturauswertung in Unterkapitel 4.1 – und dementsprechend auch die getroffene Auswahl – auf statische, tiefendaten- und merkmalsbasierte Ansätze. Dies schließt die Eignung anderer Merkmalsextraktionsverfahren selbstverständlich nicht aus. Mögliche Anknüpfungspunkte für weiterführende Arbeiten, die den Rahmen dieser Arbeit überstiegen hätten, sind unter anderem:

- 3D-Local-Binary-Patterns (3D-LBP) (z.B. [Sandbach et al., 2012b], [Bayramoglu et al., 2013])
- Kombinierte Verfahren, die zusätzlich Farb- oder Texturmerkmale extrahieren (z.B. Histogramme orientierter Gradienten [Dalal und Triggs, 2005], 2D-LBP [Ahonen et al., 2006])
- Extraktionsansätze für dynamische Merkmale (z.B. [He et al., 2009])

Feedbackableitung

In Kapitel 5 wurde das in dieser Arbeit entwickelte Verfahren zur Erzeugung von globalem und regionenbezogenem Feedback vorgestellt und evaluiert. Wesentliche Grundlage des Verfahrens sind die sogenannten paarweisen Ähnlichkeiten, die sich aus einem trainierten Random-Forest (RF) ableiten lassen (siehe Abschn. 5.2.1 und 5.2.2). Die Vorgehensweise für die Erzeugung von globalem und lokalem Feedback stimmt dabei weitestgehend überein. Daher ist, neben einem globalen RF, für jede lokale Region

ein gesonderter RF auf Basis der regionenspezifischen Merkmalsdeskriptoren zu trainieren.

Das Training des RF erfolgt unter Einsatz der Trainingsobservationen aller Übungsklassen. Für die Feedbackableitung werden anschließend nur die Trainingsobservationen der auszuführenden Übung betrachtet. Die zwischen der Patientenobservation und jeder klassenidentischen Trainingsobservation ermittelten paarweisen Ähnlichkeiten sind ein Maß für die Nähe der Observationen innerhalb des Merkmalsraumes. Der RF dient zugleich als vorgeschaltetes Merkmalsselektionsverfahren (siehe Abschn. 5.1.4). Zur Ableitung des Feedbacks werden die beschriebenen paarweisen Ähnlichkeiten den trainingsdateninternen paarweisen Ähnlichkeiten gegenübergestellt. Nähere Informationen dazu sind in den Abschnitten 5.2.3 und 5.2.4 beschrieben. Das finale Feedback umfasst sechs diskrete Feedbackstufen, die zur globalen bzw. regionenbezogenen Bewertung der Übungsausführung zugeordnet werden können.

Die Zwischenschritte und -resultate des Feedbackverfahrens wurden im Rahmen verschiedener quantitativer und qualitativer Evaluationsszenarien analysiert (Abschn. 5.3.2, 5.3.3 und 5.5.2). Für einen Ausblick lassen sich im Wesentlichen zwei Anknüpfungspunkte formulieren:

- Vergleicht man die trainingsdateninternen paarweisen Ähnlichkeiten, welche der globalen bzw. lokalen Feedbackerzeugung zu Grunde liegen, zeigt sich, dass diese im lokalen Fall die annotierte Struktur der Observationen schlechter repräsentieren. Um die Qualität des lokalen Feedbacks zu erhöhen, ist eine Anpassung und Erweiterung der lokalen Merkmalsextraktionsverfahren sinnvoll.

- Der Vorteil des vorgestellten Verfahrens ist, dass es das Feedback direkt aus den Trainingsdaten ableitet. Ein vorgeschaltetes und zeitaufwändiges manuelles Annotieren der Ground-Truth über die Übungsklasse hinaus ist nicht erforderlich. Um bei Bedarf Expertenwissen einbeziehen zu können, wäre eine Weiterentwicklung des Verfahrens um eine Art interaktives Reinforcement-Learning sinnvoll. In einem solchen Szenario könnte der Logopäde während einer Übungssitzung optional eine unmittelbare Bewertung des gegebenen Feedbacks vornehmen, die das zukünftige Verhalten des Programms beinflusst.

Visualisierung und Prototyp

Um die Ergebnisse des entwickelten Verfahrens zu veranschaulichen, wurde ein Prototyp mit grafischer Nutzeroberfläche implementiert. Dieser umfasst fünf Feedbackelemente, aus denen sich unter anderem eine globale und fünf regionenbezogene Bewer-

tungen der Übungsausführung auslesen lassen. Nähere Details wurden im Unterkapitel 5.4 beschrieben.

Der Prototyp wurde in Matlab umgesetzt und dient in erster Linie der Demonstration des resultierenden Feedbacks. Für eine laufzeitoptimierte Version ist die Portierung in eine andere Programmiersprache (z.B. C++) sinnvoll.

A. Anhang: Tiefergehende Grundlagen und Erläuterungen

In den folgenden Unterkapiteln finden sich ergänzende Informationen zu den Themenkomplexen der Hauptarbeit.

A.1. Grundlagen der Fazialisparese

Die **periphere Gesichtslähmung** ist auf eine Läsion des Gesichtsnervs (*med.* Fazialisnerv) zurückzuführen. Dieser ist der siebte von zwölf Hirnnerven und entspringt im Hirnstamm [Schmidt und Thews, 1993]. Der Fazialisnerv teilt sich in fünf Endäste auf, die unterschiedliche Muskeln des Gesichts motorisch innervieren. Ihre Namensgebung basiert auf den lateinischen Bezeichnungen der Gesichtsbereiche, in die sie münden: Schläfe (lat. *tempus*), Jochbein (lat. *Os zygomaticum*), Backe (lat. *bucca*), Unterkiefer (lat. *mandibula*), Hals (lat. *collum*). Die Abkürzungen *R.* und *Rr.* stehen für die lateinischen Übersetzungen ramus (dt. *Nervenast*, Sg.) und rami (dt. *Nervenäste*, Pl.).

Im Folgenden werden die von den Endästen innervierten Muskeln und die daraus entstehenden Gesichtsbewegungen im Detail aufgeführt ([Berghaus et al., 1996], [Graumann und Sasse, 2004], [Probst et al., 2008], [Waldeyer und Mayet, 1986]). Dabei besteht kein Anspruch auf Vollständigkeit, da die Übersicht auf die für die Aufgabenstellung relevanten Muskelgruppen reduziert ist. Angelehnt an die medizinische Fachliteratur und zur Sicherstellung einer besseren Lesbarkeit wird für den lateinischen Begriff *musculus* (dt. *Muskel*) die Abkürzung *M.* verwendet.

Die *Rr. temporales* innervieren den M. frontalis und den M. orbicularis oculi. Ersterer verläuft entlang der Stirn und bewirkt bei Kontraktion das Stirnrunzeln. Letzerer ist der sogenannte Augenringmuskel, der dem Schluss der Lidspalte dient. Beide gemeinsam wirken als Heber der Braue.

Die *Rr. zygomatici* innervieren ebenfalls den M. orbicularis oculi, sowie den M. zygomaticus major und den M. zygomaticus minor. Die beiden am Jochbein ansetzenden Muskeln ziehen bei ihrer Kontraktion die Nasenlippenfurche und den Mundwinkel

nach oben. Die *Rr. buccales* münden in den M. buccinator, der die Wangen gegen die Zahnreihen drückt und in den M. orbicularis oris - den Ringmuskel des Mundes - der für das Zusammenziehen der Mundöffnung (z.B. schließen, Kussmund) verantwortlich ist. Der innere, an das Lippenrot angrenzende, Teil wird als pars labialis bezeichnet, der äußere Teil als pars marginalis. Kontraktion des pars labialis führt zum Nachinnenstülpen der Lippen (Rot verschwindet). Eine Kontraktion des pars marginalis stülpt die Lippen nach aussen und mehr Lippenrot wird sichtbar. Ebenfalls innerviert wird der M. levator labii superioris, der Oberlippenheber.

Der *R. marginalis mandibulae* führt unter anderem motorische Fasern für den M. depressor labii inferiores, der die Unterlippe nach seitlich unten zieht.

Der *R. colli* innerviert das Platysma, einen Muskel, der sich über die vordere Halsseite erstreckt, und vor allem bei Schreckreaktionen aktiviert wird. Es zieht bei Kontraktion die Mundwinkel, Unterlippe und den Unterkiefer nach unten.

Die **zentrale Fazialisparese** ist mit einer Schädigung im Hirn verbunden und kann daher mit weiteren krankheitsbedingten Beeinträchtigungen, beispielsweise Lähmungen von Körperteilen, einhergehen. Dies ist jedoch nicht zwingend. Daher ist es notwendig, die zentrale von der peripheren Gesichtslähmung abzugrenzen, um eine korrekte - und im Falle eines Schlaganfalls umgehende - medizinische Versorgung gewährleisten zu können. Dazu überprüft man die motorische Funktion der Mimik anhand der folgenden Übungen [Berghaus et al., 1996]:

Stirn runzeln - Augen schließen - Wangen aufblasen - Mund spitzen - Pfeifen - Zähne zeigen

Hinweise auf eine periphere Fazialisparese sind die mangelnde Fähigkeit die Stirn zu runzeln und eine Vertikaldrehung des Augapfels nach oben beim Versuch des Lidschlusses. Ersteres basiert auf der unterschiedlichen Informationsversorgung der Gesichtsmuskulatur von oberer und unterer Gesichtshälfte und ist in Abbildung A.1 schematisch dargestellt.

A.2. Fitten einer Regressionsebene

Das Fitten einer Ebene in diskrete 3D-Punkte $p_i = (x_i, y_i, z_i)$, mit $i = 1, ..., n$, basiert in dieser Arbeit auf der Funktion plane_fit.m von Kevin Moerman und wurde an die Anforderungen dieser Arbeit angepasst. Die Implementierung von Kevin Moerman ist über den *Fileexchange* von Mathworks verfügbar[1].

Im ersten Schritt der Ebenenregion wird der Zentroid P der Punktwolke p_i bestimmt

[1] `http://de.mathworks.com/matlabcentral/fileexchange/22042-plane-fit/content/ /plane_fit.m`; Letzter Zugriff am 04.03.2015

und die Punktwolke so verschoben, dass P im Koordinatenursprung liegt. Dies entspricht einer Mittelwertbereinigung der Punktkoordinaten. Diese werden dann in eine $n \times 3$ dimensionale Matrix $\mathbf{M}$ zusammengefasst, auf welche anschließend eine Singulärwertzerlegung angewendet wird:

$$\mathbf{USV} = \mathbf{M}. \tag{A.1}$$

Der dritte Spaltenvektor von $\mathbf{V}$ entspricht dem Normalenvektor N der gefitteten Ebene, welcher zur Ermittlung der Ebenengleichung Einsatz findet. Für die Elemente der Koordinatenform

$$Ax_i + By_i + Cz_{fit} = D, \tag{A.2}$$

gilt $A = N(1)$, $B = N(2)$ und $C = N(3)$. Der Paramter D ergibt sich aus $\langle P, N_0 \rangle$, wobei N_0 der Normaleneinheitsvektor ist. Durch Umformen der Koordinatengleichung lässt sich der z-Wert der gefitteten Ebene z_{fit} berechnen.

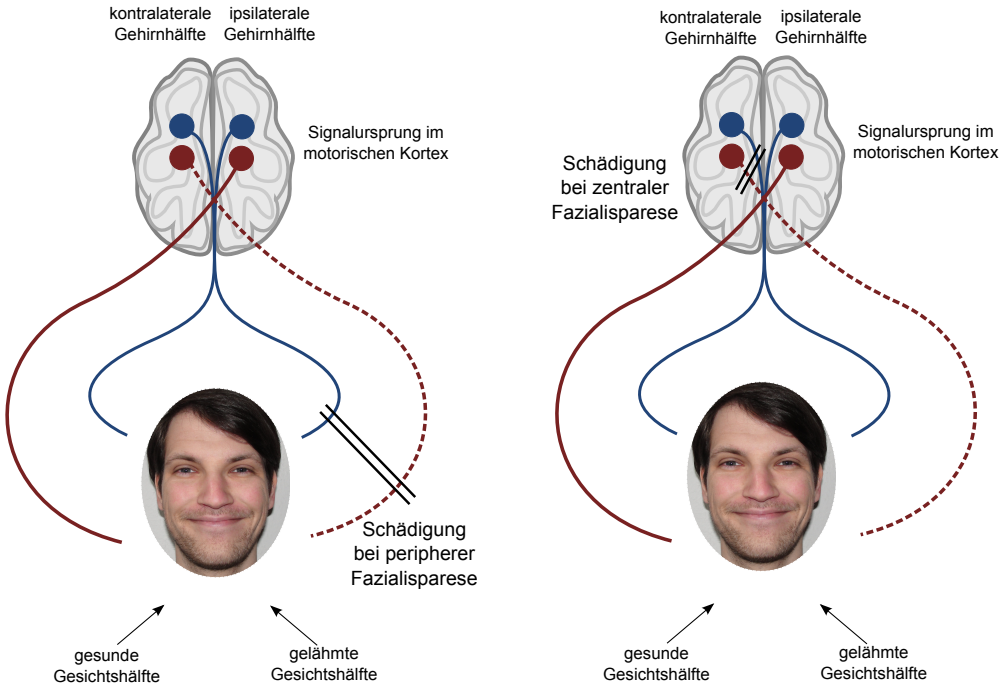


Abbildung A.1.: Versorgung der Gesichtsmuskulatur mit Informationen aus dem motorischen Kortex. Im Bereich der oberen Gesichtsmuskulatur wird eine Gesichtshälfte von beiden Gehirnhälften mit Informationen versorgt - also von der gleichseitigen (ipsilateralen) und der gegenüberliegenden (kontralateralen) Hirnhälfte. Die untere Gesichtsmuskulatur wird nur von der kontralateralen Gehirnhälfte versorgt. Im Falle einer peripheren Fazialisparese (Abb. (a)) führt eine Läsion des Fazialisnervs zu einer Versorgungsunterbrechung der gesamten Muskulatur der betroffenen Seite. Bei der zentralen Fazialisparese (Abb. (b)) liegt der Läsionsort im Gehirn und zwar noch vor der Vereinigung der Informationsflüsse der beiden Gehirnhälften. So wird die Signalleitung zur unteren Gesichtsmuskulatur in die betroffene Gesichtshälfte unterbrochen, die obere Muskulatur wird jedoch durch die ipsilaterale Gehirnhälfte mitversorgt. Das Runzeln der Stirn ist weiterhin möglich.

B. Anhang: Ergänzende Tabellen und Abbildungen

B.1. Therapeutische Übungen

Sechzehn therapeutische Fazialisübungen (drei weitere ergeben sich aus der achsengespiegelten Durchführung der Übungen *WangeLi*, *BoxenLi* und *ZungeLi*).

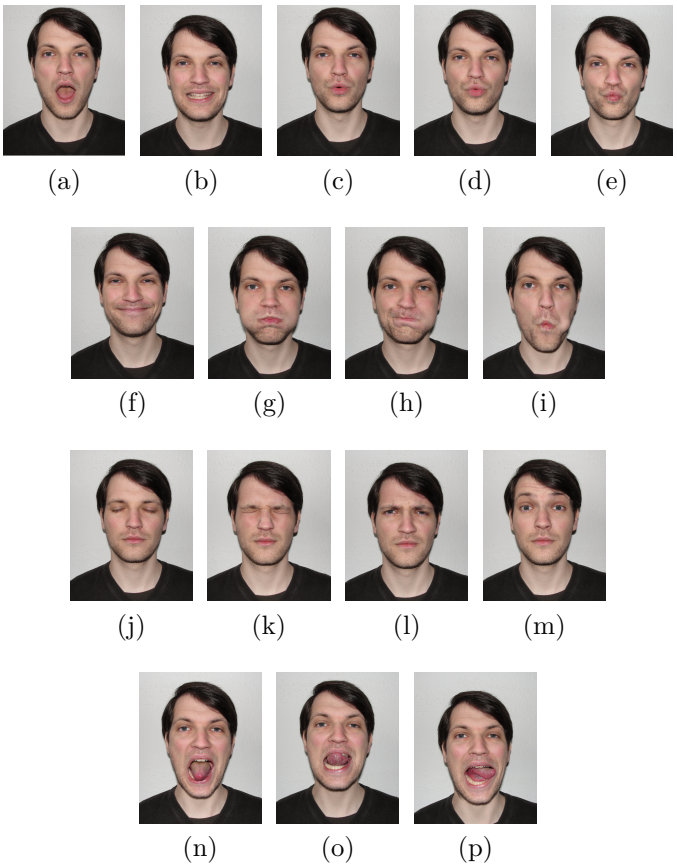


Abbildung B.1.: **(a)**–**(d)** Vokalübungen *AForm*, *IForm*, *OForm*, *UForm*. **(e)** Übung *Kuss*. **(f)** Mundübung *Breit*. **(g)**–**(i)** Wangenübungen *Wangen*, *WangeLi*, *BoxenLi*. **(j)**–**(m)** Augenübungen *Augen*, *Augenpressen*, *Brauen*, *Stirn*. **(n)**–**(p)** Zungenübungen *Zunge*, *ZungeO*, *ZungeLi*.

B.2. Literaturauswertung

In den folgenden Tabellen finden sich detaillierte Informationen zu den einzelnen recherchierten Publikationen des Anwendungsszenarios (Abschn. 2.1.1) und der Merkmalsextraktion (Abschn. 4.1).

B.2.1. Literaturübersicht: Anwendungsszenario (27 Publikationen)

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
1	[Brach und VanSwearingen, 1999]	D, T, *A*	-	Medizinische Fallstudie zur Rehabilitation einer FP-Patientin (Diagnose und Therapie). Beobachtungszeitraum über 13 Monate. Übungsfeedback mittels EMG und Spiegeln.
2	[Gebhard et al., 2000]	D, T, *A*	2D, EB	Ziel ist die Quantifizierung der Asymmetrie. Wissensbasierter Ansatz zur automatisierten Gesichts- und Landmarkenlokalisierung. Relevante LM sind die äußeren Augen- und Mundwinkel. Mit Hilfe von orientierbaren 2D-Gaußfiltern werden an allen vier LM Signaturen extrahiert. Die Korrelation der Signaturen der linken und rechten GH dient als Maß der Asymmetrie.

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
3	[Gebhard et al., 2001]	D, *A*	2D, V	Vorstellung eines Systems zur Gradierung des Lähmungsgrades auf Basis eines Nächste-Nachbarn-Klassifikators. Die vier Bewertungsklassen orientieren sich am sechsstufigen H-B-G. Zum Zweck der Merkmalsextraktion werden zwischen aufeinanderfolgenden Frames Intensitätsdifferenzbilder und der optische Fluss bestimmt. Von beiden werden im Anschluss die Differenzen zwischen der linken und rechten GH ermittelt.
4	[Wachtman et al., 2001]	-	2D, V	Vergleich eines automatisierten Verfahrens zur Landmarkenverfolgung (Lukas-Kanade-Tracker) mit einem etablierten manuellen Trackingverfahren (MS-RA). Ein manuelles Positionieren der Landmarken im ersten Frame bleibt erforderlich.
5	[Cronin und Steenerson, 2003]	D, T	-	Retrospektive, medizinische Fallanalysen. Auswertung der Rehabilitationsverläufe von 24 Patienten (+ Kontrollgruppe). Feedback mittels EMG und Spiegeln.
6	[Nakamura et al., 2003]	D,(T), *A*	2D, EB	Medizinische Studie mit 27 Patienten (+ Kontrollgruppe) über die Synkinese-Prävention im Rahmen der FP-Rehabilitation. Ablauf der Therapie klassisch ohne Computerunterstützung (Biofeedback mit Hilfe von Spiegeln). Diagnostizierung einer möglichen Synkinese durch Vergleich der linken und rechten vertikalen Augenöffnung (Abstandsmessung in Photoshop).

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
7	[S. Wang und Qi, 2005]	D, *A*	2D, EB	Bestimmung der Asymmetrie zwischen linker und rechter GH zur Quantifizierung des Lähmungsgrades. Die Merkmalsextraktion erfolgt sowohl auf Basis der Differenzbilder zwischen der linken und rechten GH (*Dface*), als auch dem Vergleich von vier mimischen Übungen mit dem Neutralgesicht (*Pface*). Vergleich des Resultats mit dem H-B-G. Helmkamera zur Bildaufnahme, um einheitliche Orientierung des Gesichts zu gewährleisten. Manuelles Setzen von vier LM im ersten Frame als Ausgangspunkt für automatisiertes LM-Tracking.
8	[He et al., 2007]	D, *A*	2D, V	Ziel ist die automatisierte Klassifikation des H-B-G. Dazu wird für insgesamt fünf Mimikübungen jeweils ein RGB-Netz trainiert und im Anschluss aus den fünf Ausgabewerten ein allgemeiner H-B-G ermittelt. Der Bewegungverlauf zwischen Neutralgesicht und Übungsextremum wird dabei mittels Optical Flow geschätzt und die Differenz zwischen linker und rechter GH extrahiert, wobei der vertikale Bewegungsanteil stärker gewichtet wird. Automatisierte Lokalisierung von Gesicht und LM mit strengen Vorgaben (homogener Hintergrund, Frontalansicht).
9	[Dong et al., 2008]	D, *A*	2D, EB	Wissensbasierter Ansatz zur Gesichtsdetektion. Ableitung von relevanten LM auf Basis von Salient-Points und Kantendetektoren. Analyse von fünf Mimikübungen. Das Distanzverhältnis zwischen gesunder und beeinträchtigter GH dient als Maß für Asymmetrie und Lähmungsgrad. Kein Mapping auf medizinische Gradierungsskala.

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
10	[He et al., 2008]	D, T, *A*	2D, V	Ziel ist die Modellierung, Analyse und Visualisierung der Asymmetrie zwischen linken und rechten Gesichtsregionen über den gesamten zeitlichen Verlauf einer Übungsausführung hinweg. Dazu werden Local Binary Patterns extrahiert und mittels Local Linear Embedding in einen niedrigdimensionalen Raum überführt. Als Distanzmaß dient die modifizierte mittlere Hausdorff-Distanz.
11	[He et al., 2009]	D, *A*	2D, V	Klassifikationsbasierte Bestimmung von fünf regionalen und einem globalen H-B-G. Raum- und zeitbezogene Extraktion von Local-Binary-Patterns. Resistor-Average-Distance-Berechnung zwischen Merkmalsvektoren der linken und rechten GH. Die Abstandsmerkmale werden einem Klassifikator übergeben (SVM, RBF-NN, kNN).
12	[Delannoy und Ward, 2010]	D, *A*	2D, EB	Konzeptionelle Überlegungen für ein Therapieszenario. Die experimentelle Auswertung bezieht sich jedoch auf die Bestimmung des H-B-G. Laterale Verfolgung der Mundwinkel zwischen Neutralgesicht und Lächeln. Extrahierter Distanzwert wird regelbasiert auf H-B-G gemappt. Der Originaldatensatz umfasst Einzelbilder von gesunden Personen, der FP-Datensatz wurde mittels Active-Appearance-Models synthetisiert.
13	[Liu et al., 2010]	D, *A*	2D, EB	Manuelles setzen einer Stirn-LM direkt auf der Haut zur Unterteilung des Gesichts in eine linke und rechte GH. Binäre Segmentierung um lokale Areale im Gesicht (Nasenlöcher, Mund, Stirnfalten) freizustellen. Asymmetrie wird aus den Flächenrelationen der linken und rechten Regionenhälften geschätzt. Kein Mapping auf medizinische Gradierungsskala.

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
14	[Kihara et al., 2011]	D, *A*	2D, EB	Distanzbasierte Evaluierung und Verfolgung der LM-Positionen für verschiedene Gesichtsausdrücke. Schaubildbasierte Gegenüberstellung der extrahierten Distanzewerte von gesunden Personen und Patienten. Kein Mapping auf medizinische Gradierungsskala.
15	[Nöth et al., 2011]	D, T	2D, 3D	Der Fokus der Veröffentlichung liegt auf der Evaluation von Sprechstörungen (Dysarthrie). Die Veröffentlichung gliedert sich in zwei Teile. Der erste Teil beschränkt sich methodisch und experimentell auf Audiodaten. Im zweiten Teil wird ein Konzept für ein telemedizinisches Diagnose- und Therapiesystem vorgestellt, bei welchem auch die visuelle Auswertung der Übungen thematisiert wird. Für konkrete methodische Ansätze wird auf [Gebhard et al., 2001] verwiesen.
16	[Jayatilake et al., 2012]	T, *A*	EMG	Fallstudie (1 Patientin) zum Einsatz einer sogenannten Robot-Mask für die FP-Rehabilitation. Die Schnittstelle wird am Kopf getragen. Mit Hilfe von Oberflächenelektroden werden bioelektrische Signale der gesunden GH aufgezeichnet und im Anschluss, mittels Aktuatoren und Klebebefestigungen, auf die andere GH übertragen um dort eine Bewegung auszulösen. Die Fallstudie untersucht den Tragekomfort, die Latenz der Übertragung, sowie den Einfluss auf die Symmetrie der Mimikbewegungen. Verfahren teilautomatisiert, da zu Beginn Anbringung der Elektroden und Klebebefestigungen notwendig ist.

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
17	[Hadlock und Urban, 2012]	D	2D, EB	VÖ mit eher medizinischem Hintergrund. Vorstellung der Software FACE, welche zur Messung von Distanzen zwischen LM dient und die, bis dahin eingesetzte, manuelle Messung in Photoshop ersetzen soll. Eine manuelle Positionierung der LM im Bild bleibt erforderlich. Die Evaluation umfasst die Gegenüberstellung der FACE- und Photoshop-Messergebnisse.
18	[Kleiss et al., 2013]	D, *A*	2D, EB	VÖ mit eher medizinischem Hintergrund. Ziel ist die Evaluation der okularen Synkinese auf Basis der einzelnen Augenöffnungen (AÖ), sowie der AÖ-Symmetrie. Zu diesem Zweck wird der Einsatz der FACE-Software und der modifizierten Glasgow-Facial-Palsy-Scale untersucht.
19	[L. Y. Lin, 2013]	D, *A*	3D, EB	Doktorarbeit mit dem Ziel der Entwicklung eines FP-Diagnosesystems unter Verwendung von 3D-Daten (triangulierte Polygonnetze). In der experimentellen Auswertung erfolgt keine Gradierung, sondern lediglich eine binäre Entscheidung (FP ja/nein) auf Basis von Noise-Injected-ANNs. Extrahierte Merkmale sind die Gauß'sche Krümmung, der Shape-Index, sowie die geometrische Distanz zwischen den registrierten linken und rechten GH. Anhand von Grenzwerten werden für alle Merkmale Asymmetrie-Indizes bestimmt und dem ANN übergeben.

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
20	[Tasneem et al., 2014]	T	2D, EB	Vorstellung eines interaktiven Spiels, bei welchem ein Ball auf einer Wippe balanciert werden soll. Die Neigung der Wippe wird durch den Grad der Augenöffnung bestimmt. Dieser ergibt sich aus dem Abstand zwischen oberem und unterem Augenlid. Die Lokalisierung der Augenregion erfolgt automatisiert unter Einsatz der OpenCV.
21	[Y.-X. Wang et al., 2014]	T	3D, V	Vorstellung eines interaktiven Spiels. Der Spieler sammelt Punkte, indem er virtuelles Essen anbeißt oder mit der Zungenspitze berührt. Die Umsetzung basiert auf der Kinect und der Microsoft Face Tracking SDK. Letztere enthält Funktionen zur Lokalisierung der Mundregion und Erkennung der Beißbewegung. Die Experimente konzentrieren sich auf die Erkennungsrate der Beißbewegung (100%) und die Genauigkeit der Zungenspitzenlokalisierung. Das Spielkonzept wird nicht evaluiert.
22	[Park und Oh, 2015]	(T)	2D, EB	Rekonstruktion eines nutzerspezifischen Bilinear-Shape-Model aus RGB-Bildern einer Webcam. Das 3D-Modell wird auf Basis der Nutzeridentität und des Gesichtsausdrucks rekonstruiert und mit einem Zielmodell abgeglichen, um zu erkennen, ob die ausgeführte Übung der Zielübung entspricht. Möglicher Therapieeinsatz des Verfahrens wird erwähnt, es ist jedoch keine Methode für das Mappen der Modellparameter auf Patientenfeedback beschrieben. Keine experimentelle Evaluation des Verfahrens.

Tabelle B.1.: Veröffentlichungen aus dem Themenbereich der Fazialisparese-Rehabilitation. In der dritten Spalte ist vermerkt, ob die beschriebenen Verfahren der Diagnose oder der Therapie (D, T) dienen und Aspekte der Asymmetrieevaluation enthalten (*A*). Ein grau unterlegter Eintrag kennzeichnet vollautomatisierte Verfahren. Die vierte Spalte enthält Informationen zur Datenbasis und Merkmalsextraktion, bezogen auf räumliche und zeitliche Eigenschaften (2D, 3D; Einzelbilder (EB), Videos (V)). Abschließend findet sich eine Kurzfassung der Ansätze. Verwendete Abkürzungen innerhalb der Beschreibung: *LM* Landmarken, *GH* Gesichtshälfte, *H-B-G* House-Brackmann-Grad, *FP* Fazialisparese.

Nr.	Publikation	Szenario	Daten	Beschreibung
23	[Gaber et al., 2015]	D, *A*	3D, EB	Ziel ist die Quantifizierung des FP-Grades. Verwendung der Kinect V2, sowie der Kinect for Windows SDK 2.0. Extrahierte Merkmale sind Distanzen zwischen Landmarken (in 3D), die Schiefe des Mundes und die Fläche der Augenöffnung, angenähert mittels einer Ellipse. Die Evaluierung basiert ausschließlich auf einem Datensatz von gesunden Personen.
24	[T. Wang et al., 2015]	D, *A*	2D, V	Klassifikationsbasiertes, automatisiertes Verfahren zur Ermittlung des H-B-G (statisch und dynamisch). Bei der statischen Merkmalsextraktion wird das Gesicht in acht Regionen unterteilt. Zwischen gegenüberliegenden Regionen werden Differenzbilder berechnet, welche zur Local-Binary-Pattern-Extraktion dienen. Der dynamische Kennwert beschreibt die zeitliche Abweichung zwischen beiden Gesichtshälften bei dem Übergang vom Neutralgesicht zum Maximum der auszuführenden Mimik.
25	[Kim et al., 2015]	D, *A*	2D, V	Vorstellung einer Smartphone-Anwendung zur Feststellung einer möglichen Fazialisparese (binäre Klassifikation). Punkt- und achsenbasierte Distanzextraktion. LM-Lokalisierung nach [Asthana et al., 2014].
26	[Haase et al., 2015]	D	2D, EB	Vorstellung einer Methode für die objektive Bewertung des Fazialisparesegrades, basierend auf Active-Appearance-Models (AAM) und der automatisierten Detektion von Action-Units (AU).
27	[Modersohn und Denzler, 2016]	D	2D, EB	Vorstellung einer Methode zur automatisierten Klassifikation der House-Brackmann und Stennert Index Diagnosestufen. Zugrunde liegende Konzepte sind AAMs, AUs, Landmarkendistanzen und Random-Decision-Forests.

B.2.2. Literaturvergleich: Hauptpunktverschiebung

Die nachstehende Tabelle B.2 vergleicht die anhand der Kalibrierung geschätzte Hauptpunktverschiebung $\Delta\mathbf{p}_0 = [\Delta x, \Delta y]$ für das RGB- und das IR-Bild mit in der Literatur aufgeführten Beispielwerten.

Tabelle B.2.: Hauptpunktverschiebung $\Delta\mathbf{p}_0 = [\Delta x, \Delta y]$. Die Verschiebung ist in Pixeln angegeben.

	RGB		IR	
	Δx	Δy	Δx	Δy
Eigene Ergebnisse	**-12**	**26**	**8**	**2**
[Khoshelham und Elberink, 2012]	4	-35	-7	-4
[Herrera et al., 2011]	-1	22	3	-9
[Menna et al., 2011]	6	-6	5	5
[Smisek et al., 2013]	-3	-2	-4	8

B.2.3. Literaturübersicht: Merkmalsextraktion (28 Publikationen)

Tabelle B.3.: Übersicht über die ausgewerteten Publikationen zur Merkmalsextraktion. Grundsätzlich werden die Ansätze in distanz- (D), patchbasierte (P) oder globale (G) Verfahren unterteilt. Eine graue Einfärbung in der vierten Spalte (Ziel) kennzeichnet Publikationen, in denen der BU-3DFE-Datensatz zur Evaluierung eingesetzt wird. Eine graue Einfärbung in der fünften Spalte weist auf Verfahren hin, die eine automatisierte Landmarkenlokalisierung integriert haben oder keine Landmarkenlokalisierung benötigen. Verwendete Abkürzungen: *LM* Landmarken, *FER* facial expression recognition, *AUD* action unit detection, *EVD* emotion valence detection, *GR* gender recognition.

Nr.	Publikation	D,P,G	Ziel	LM	Extrahierte Merkmale
1	[J. Wang et al., 2006]	P	FER	47	Aus einer Auswahl von 12 kategorischen Krümmungstypen wird jedem Vertex des Gesichts ein Typ zugeordnet. Dann wird das Gesicht in sieben Regionen unterteilt und für jede Region ein Histogramm über die Verteilung der Krümmungstypen ermittelt.
2	[Soyel und Demirel, 2007]	D	FER	11	Extraktion von sechs 3D-Distanzen zwischen den LM.

Tabelle B.3.: Übersicht über die ausgewerteten Publikationen zur Merkmalsextraktion. Grundsätzlich werden die Ansätze in distanz- (D), patchbasierte (P) oder globale (G) Verfahren unterteilt. Eine graue Einfärbung in der vierten Spalte (Ziel) kennzeichnet Publikationen, in denen der BU-3DFE-Datensatz zur Evaluierung eingesetzt wird. Eine graue Einfärbung in der fünften Spalte weist auf Verfahren hin, die eine automatisierte Landmarkenlokalisierung integriert haben oder keine Landmarkenlokalisierung benötigen. Verwendete Abkürzungen: *LM* Landmarken, *FER* facial expression recognition, *AUD* action unit detection, *EVD* emotion valence detection, *GR* gender recognition.

Nr.	Publikation	D,P,G	Ziel	LM	Extrahierte Merkmale
3	[P. Wang et al., 2007]	D,P	FER	58	Extraktion von 2D- und 3D-Merkmalen, sowie Verbindung beider Modalitäten über Bildmomente (*engl.* image moments). In 2D werden neun Distanzen zwischen den LM und die Größe von 28 Regionen extrahiert. In 3D wird für 21 Regionen je ein Histogramm über vier Krümmungstypen erstellt. Die automatisierte Lokalisierung der LM erfolgt im Farbbild.
4	[H. Tang und Thomas S. Huang, 2008]	D	FER	83	Extraktion von 24 3D-Distanzen zwischen den LM.
5	[Soyel und Demirel, 2008]	D	FER	24	Extraktion von sechs 3D-Distanzen. Im Vergleich zu [Soyel und Demirel, 2007] werden mehr LM verwendet und die resultierenden Distanzen aus bis zu drei anderen gemittelt.
6	[H. Tang und Thomas S Huang, 2008]	D	FER	83	Extraktion von 24 Distanzen und 24 Richtungsvektoren (*engl.* slope features) zwischen den LM.
7	[Srivastava und Roy, 2009]	D	FER	83	Extraktion der Magnitude und der Richtung des Versatzes der 83 LM im Vergleich zum Normalgesicht.
8	[Maalej et al., 2010]	P	FER	24	Bestimmung der Deformation von kleinen Patches im radialen Umfeld von LM.
9	[Savran et al., 2010]	G	AUD	-	Vergleich 2D und 3D Merkmale (Gabor Wavelets und Krümmungsanalyse), sowie die Kombination über Konkatenierung. Globaler datengetriebener Ansatz (Input ist 96×96 Pixel großes Bild).
10	[X. Li et al., 2010]	D	FER	35	Distanz-, Slope- und Winkelmerkmale zwischen LM.

Tabelle B.3.: Übersicht über die ausgewerteten Publikationen zur Merkmalsextraktion. Grundsätzlich werden die Ansätze in distanz- (D), patchbasierte (P) oder globale (G) Verfahren unterteilt. Eine graue Einfärbung in der vierten Spalte (Ziel) kennzeichnet Publikationen, in denen der BU-3DFE-Datensatz zur Evaluierung eingesetzt wird. Eine graue Einfärbung in der fünften Spalte weist auf Verfahren hin, die eine automatisierte Landmarkenlokalisierung integriert haben oder keine Landmarkenlokalisierung benötigen. Verwendete Abkürzungen: *LM* Landmarken, *FER* facial expression recognition, *AUD* action unit detection, *EVD* emotion valence detection, *GR* gender recognition.

Nr.	**Publikation**	**D,P,G**	**Ziel**	**LM**	**Extrahierte Merkmale**
11	[Yun und Guan, 2010]	G	FER	-	Globaler Ansatz mit 3D Gabor Filtern.
12	[Maalej et al., 2011]	P	FER	68	Deformationsbestimmung von kleinen Patches im radialen Umfeld von LM.
13	[Lemaire et al., 2011]	P	FER	19	Berechnung der mittleren euklidische Distanz zwischen registrierten Test- und Referenzdaten für zehn Regionen des Gesichts. Automatisierte Lokalisierung der LM durch Verwendung eines statistischen Models für die Verteilung der Landmarken.
14	[Vretos et al., 2011]	G	FER	-	Extraktion von Zernike-Momenten. Es werden keine Landmarken benötigt.
15	[Berretti et al., 2011]	P	FER	9	Extraktion von SIFT-Deskriptoren aus der Umgebung von 79 Punkten der Gesichtsoberfläche. Die 79 Punkte werden von 9 automatisiert lokalisierten LM abgeleitet.
16	[C. Li und Soares, 2011]	D	FER	5	Extraktion von Distanz- und Winkelmerkmalen zwischen LM.
17	[Sandbach et al., 2012b]	P	AUD	6	Extraktion von LNBPs (Local Normal Binary Patterns). Dann wird das Gesicht in 10×10 Blöcke unterteilt und für jeden Block ein Histogramm erstellt.

Tabelle B.3.: Übersicht über die ausgewerteten Publikationen zur Merkmalsextraktion. Grundsätzlich werden die Ansätze in distanz- (D), patchbasierte (P) oder globale (G) Verfahren unterteilt. Eine graue Einfärbung in der vierten Spalte (Ziel) kennzeichnet Publikationen, in denen der BU-3DFE-Datensatz zur Evaluierung eingesetzt wird. Eine graue Einfärbung in der fünften Spalte weist auf Verfahren hin, die eine automatisierte Landmarkenlokalisierung integriert haben oder keine Landmarkenlokalisierung benötigen. Verwendete Abkürzungen: *LM* Landmarken, *FER* facial expression recognition, *AUD* action unit detection, *EVD* emotion valence detection, *GR* gender recognition.

Nr.	Publikation	D,P,G	Ziel	LM	Extrahierte Merkmale
18	[Sandbach et al., 2012a]	P	AUD	6	Basierend auf der Repräsentation des Gesichts als Tiefenbild und als APDI (Azimuthal Projection Distance Image) werden sieben verschiedene Binärmuster-Merkmalsdeskriptoren extrahiert. Das Gesicht wird in 10×10 Blöcke unterteilt und für jeden Block wird je ein Histogramm für die sieben Merkmalsdeskriptoren berechnet. Sechs LM werden zur Ausrichtung der 3D-Gesichter benötigt.
19	[Y. Wang und Ma, 2012]	P	FER	?	Extrahiert werden die Tiefenwerte der einzelnen Vertizes, sowie die x-, y- und z-Werte der Normalenvektoren. Dann wird das Gesicht in 21 Regionen unterteilt und für jede Region vier Histogramme für die einzelnen Merkmalsarten erstellt. Aufteilung der 21 Regionen entspricht nicht einem regelmäßigen Gitter. Die genaue Vorgehensweise bei der Aufteilung und die dafür benötigte Anzahl an LM wird nicht deutlich. Drei LM werden genannt - die Nasenspitze wird über die minimale Kameradistanz und Vektorauslenkung bestimmt, die beiden Augen über HSV-Werte.
20	[Rabiu et al., 2012]	D	FER	29	Extraktion von 16 Distanz- und 27 Winkelmerkmalen zwischen LM.
21	[Broadbent et al., 2012]	P	FER	-	Krümmungsanalyse durch Extraktion des Shape-Index und der Curvedness. Gesicht wird unterteilt in $n \times m$ Blöcke und pro Block ein sogenanntes Area-weighted-Histogram berechnet.

Tabelle B.3.: Übersicht über die ausgewerteten Publikationen zur Merkmalsextraktion. Grundsätzlich werden die Ansätze in distanz- (D), patchbasierte (P) oder globale (G) Verfahren unterteilt. Eine graue Einfärbung in der vierten Spalte (Ziel) kennzeichnet Publikationen, in denen der BU-3DFE-Datensatz zur Evaluierung eingesetzt wird. Eine graue Einfärbung in der fünften Spalte weist auf Verfahren hin, die eine automatisierte Landmarkenlokalisierung integriert haben oder keine Landmarkenlokalisierung benötigen. Verwendete Abkürzungen: *LM* Landmarken, *FER* facial expression recognition, *AUD* action unit detection, *EVD* emotion valence detection, *GR* gender recognition.

Nr.	Publikation	D,P,G	Ziel	LM	Extrahierte Merkmale
22	[Bayramoglu et al., 2013]	D,P	AUD	26	Über das Gesicht wird ein 5×5 Blöcke umfassendes Gitter gelegt und aus jedem Block ein Histogramm für CS-3DLBP (Center Symmetric Local Binary Patterns) extrahiert. Zur Extraktion des RbG-descriptors (ratio based geometric descriptor) werden Verhältnisse zwischen LM-Distanzen oder Regionen berechnet und Winkel zwischen den Distanzstrecken ermittelt.
23	[Lemaire et al., 2013]	P	FER	1	Repräsentation des Gesichts durch DMCMs (Differential Mean Curvature Maps). Danach wird das Gesicht in mehrere Blöcke unterteilt und aus jedem Block ein HOG (Histogram of Oriented Gradients) extrahiert. Dem HOG ist eine Normalisierung des Gesichts (durch Anpassen der Bildseitenverhältnisse) vorgeschaltet, die ein Beschneiden der Gesichtsregion erfordert. Dazu werden für die weiteren Schritte nur die Teile des Gesichts beibehalten, die, von der Nasenspitze ausgehend, innerhalb eines Radiuses von 8 cm liegen.
24	[Savran et al., 2013]	P	EVD	-	Berechnen der mittleren Krümmung. Das Gesicht wird in 16×16 Blöcke unterteilt und pro Block ein Histogram für die mittlere Krümmung ermittelt. Die Publikation verwendet Daten mit vergleichsweise geringer Qualität, aufgenommen mit der Kinect.

Tabelle B.3.: Übersicht über die ausgewerteten Publikationen zur Merkmalsextraktion. Grundsätzlich werden die Ansätze in distanz- (D), patchbasierte (P) oder globale (G) Verfahren unterteilt. Eine graue Einfärbung in der vierten Spalte (Ziel) kennzeichnet Publikationen, in denen der BU-3DFE-Datensatz zur Evaluierung eingesetzt wird. Eine graue Einfärbung in der fünften Spalte weist auf Verfahren hin, die eine automatisierte Landmarkenlokalisierung integriert haben oder keine Landmarkenlokalisierung benötigen. Verwendete Abkürzungen: *LM* Landmarken, *FER* facial expression recognition, *AUD* action unit detection, *EVD* emotion valence detection, *GR* gender recognition.

Nr.	Publikation	D,P,G	Ziel	LM	Extrahierte Merkmale
25	[Zeng et al., 2013]	G	FER	3	Extraktion von MCI (Mean Conformal Image) und CFI (Conformal Factor Image). Automatisiert detektierte LM werden zur Normalisierung eingesetzt und umfassen die Nasenspitze, sowie linken und rechten Augeninnenwinkel.
26	[Huynh et al., 2013]	P	GR	6-25	Aufteilung des Gesichts in 8×8 Blöcke. Extraktion von G-LBP (Gradient Local Binary Patterns) aus jedem Block. Die LM werden zum Beschneiden des Gesichts benötigt.
27	[Y. Wang et al., 2013]	P	FER	-	Das Gesicht wird in 10×8 Blöcke unterteilt und für jeden Block mittels Krümmungsanalyse (Hauptkrümmungen k_1 und k_2, mittlere Krümmung und Shape-Index) und anschließender Codierung in LBPs (Local Binary Patterns) pro Krümmungstyp ein Histogramm ermittelt.
28	[Xue et al., 2014]	P	FER	3/25	Extraktion von lokalen Patches im Umfeld der LM mit anschließender Merkmalsreduktion. Unter Verwendung des Tiefenbildes und seinen Gradientenbildern werden mit Hilfe von Haar-like Merkmalen und AdaBoost drei LM automatisiert detektiert (Augenzentrum und die Nasenspitze). Basierend auf den drei LM werden die vier Augeninnenwinkel und dann 25 weitere LM abgeleitet.

B.3. Feedbackableitung

Tabelle B.4.: Zuordnung der Distanz- und Winkelmerkmalsvariablen zu den fünf Feedbackregionen.

MV	**Region**	**MV**	**Region**	**MV**	**Region**	**MV**	**Region**
δ_1	1	δ_{12}	3, 4, 5	θ_7	2	θ_{18}	4
δ_2	2	δ_{13}	3, 4, 5	θ_8	2	θ_{19}	4
δ_3	1,2	δ_{14}	3, 4, 5	θ_9	1	θ_{20}	4
δ_4	1,2	δ_{15}	3, 4, 5	θ_{10}	1	θ_{21}	5
δ_5	1,2	δ_{16}	3, 4, 5	θ_{11}	1	θ_{22}	3
δ_6	1	θ_1	1	θ_{12}	2	θ_{23}	4
δ_7	2	θ_2	2	θ_{13}	2	θ_{24}	5
δ_8	1,2	θ_3	1	θ_{14}	2	θ_{25}	3, 4, 5
δ_9	3, 4, 5	θ_4	1	θ_{15}	3	θ_{26}	3, 4, 5
δ_{10}	3, 4, 5	θ_5	1	θ_{16}	5	θ_{27}	4
δ_{11}	3, 4, 5	θ_6	2	θ_{17}	3		

Literatur

Abate, Andrea F., Michele Nappi, Daniel Riccio und Gabriele Sabatino (2007). "2D and 3D face recognition: A survey". In: *Pattern Recognition Letters* 28.14, S. 1885–1906 (siehe S. 56).

Ahonen, Timo, Abdenour Hadid und Matti Pietikainen (2006). "Face description with local binary patterns: Application to face recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12, S. 2037–2041 (siehe S. 171).

Asthana, Akshay, Stefanos Zafeiriou, Shiyang Cheng und Maja Pantic (2013). "Robust discriminative response map fitting with constrained local models". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 3444–3451. doi: 10.1109/CVPR.2013.442 (siehe S. 154).

— (2014). "Incremental face alignment in the wild". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1859–1866 (siehe S. 152, 154, 156, 171, 186).

Barth, Erhardt, Terry Caelli und Christoph Zetzsche (1993). "Image encoding, labeling, and reconstruction from differential geometry". In: *CVGIP: Graphical Models and Image Processing* 55.6, S. 428–446 (siehe S. 93, 98).

Bayramoglu, Neslihan, Guoying Zhao und Matti Pietikainen (2013). "CS-3DLBP and geometry based person independent 3D facial action unit detection". In: *IEEE International Conference on Biometrics (ICB)*, S. 1–6 (siehe S. 60–63, 171, 191).

Berghaus, Alexander, Gerhard Rettinger, Gerhard Böhme und Wolfgang Pirsig (1996). *Hals-Nasen-Ohren-Heilkunde: Nervus facialis*. Letzter URL-Zugriff: 01.03.2016. Hippokrates. url: http://www.uniklinik-ulm.de/fileadmin/Kliniken/HNO/lehre/duale_reihe_hno-b.pdf (siehe S. 17, 18, 174, 175).

Berretti, Stefano, Boulbaba Ben Amor, Mohamed Daoudi und Alberto Del Bimbo (2011). "3D facial expression recognition using SIFT descriptors of automa-

tically detected keypoints". In: *The Visual Computer* 27.11, S. 1021–1036 (siehe S. 61–63, 87, 189).

Besl, Paul J und Ramesh C Jain (1986). "Invariant surface characteristics for 3D object recognition in range images". In: *Computer Vision, Graphics, and Image Processing* 33.1, S. 33–80 (siehe S. 93, 95, 96, 98–101, 110).

Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan und Uri Shaft (1999). "When is "nearest neighbor" meaningful?" In: *Proceedings of the 7th International Conference on Database Theory (ICDT)*, S. 217–235 (siehe S. 137).

Bibliographisches Institut GmbH (2013). *Duden*. Dudenverlag. URL: `http://www.duden.de/rechtschreibung/Innervation` (siehe S. 18).

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Springer New York (siehe S. 113, 114).

Brach, Jennifer S. und Jessie M. VanSwearingen (1999). "Physical therapy for facial paralysis: a tailored treatment approach". In: *Physical Therapy* 79.4, S. 397–404 (siehe S. 1, 13, 14, 179).

Bradski, Gary und Adrian Kaehler (2008). *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc. (siehe S. 30, 32–34, 36).

Broadbent, Laurence, Khemraj Emrith, Abdul R. Farooq, Melvyn L. Smith und Lyndon N. Smith (2012). "2.5 d facial expression recognition using photometric stereo and the area weighted histogram of shape index". In: *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, S. 490–495 (siehe S. 62, 63, 190).

Camplani, Massimo und Luis Salgado (2012). "Efficient spatio-temporal hole filling strategy for kinect depth maps". In: *Three-Dimensional Image Processing (3DIP) and Applications II*. Hrsg. von SPIE Proceedings. Bd. 8290 (siehe S. 47, 49).

Cantzler, Helmut und Robert B. Fisher (2001). "Comparison of HK and SC curvature description methods". In: *Third International Conference on 3-D Digital Imaging and Modeling*, S. 285–291 (siehe S. 93).

Chang, Chih-Chung und Chih-Jen Lin (2011). "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (3). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 27:1–27:27 (siehe S. 67).

Chen, Xin und Francis Schmitt (1992). "Intrinsic surface properties from surface triangulation". In: *Proceedings of the Second European Conference on Computer Vision (ECCV)*. Springer, S. 739–743 (siehe S. 93, 94).

Chua, Chin-Seng, Feng Han und Yeong-Khing Ho (2000). "3D human face recognition using point signature". In: *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, S. 233–238 (siehe S. 64, 65, 78, 81, 110).

Chua, Chin-Seng und Ray Jarvis (1997). "Point signatures: a new representation for 3D object recognition". In: *International Journal of Computer Vision* 25.1, S. 63–85 (siehe S. 65).

Cohn, Jeffrey F., Zara Ambadar und Paul Ekman (2007). "Observer-based measurement of facial expression with the Facial Action Coding System". In: *The Handbook of Emotion Elicitation and Assessment*, S. 203–221 (siehe S. 115, 117).

Colombo, Alessandro, Claudio Cusano und Raimondo Schettini (2006). "3D face detection using curvature analysis". In: *Pattern Recognition* 39.3, S. 444–455 (siehe S. 95, 96, 110).

Cootes, Timothy F., Gareth J. Edwards und Christopher J. Taylor (2001). "Active appearance models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6, S. 681–685 (siehe S. 22, 61, 153).

Cronin, Gaye W. und Ronald Leif Steenerson (2003). "The effectiveness of neuromuscular facial retraining combined with electromyography in facial paralysis rehabilitation". In: *Otolaryngology–Head and Neck Surgery* 128.4, S. 534–538 (siehe S. 13, 180).

Cutler, Adele, D. Richard Cutler und John R. Stevens (2012). "Random forests". In: *Ensemble Machine Learning*. Springer, S. 157–175 (siehe S. 119–121, 124–126, 134, 137).

Dalal, Navneet und Bill Triggs (2005). "Histograms of oriented gradients for human detection". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 886–893 (siehe S. 171).

Danciu, Gabriel, Simona Maria Banu und Alexandru Caliman (2012). "Shadow removal in depth images morphology-based for kinect cameras". In: *16th Interna-*

tional Conference on System Theory, Control and Computing (ICSTCC), S. 1–6 (siehe S. 48–50, 52).

Delannoy, Jane Reilly und Tomas E. Ward (2010). "A preliminary investigation into the use of machine vision techniques for automating facial paralysis rehabilitation therapy". In: *Irish Signals and Systems Conference (ISSC 2010)*, S. 228–232. doi: 10.1049/cp.2010.0517 (siehe S. 14, 182).

Dittmar, Cornelia, Joachim Denzler und Horst-Michael Gross (2014). "Facial movement dysfunctions: Conceptual design of a therapy-accompanying training system". In: *Ambient Assisted Living*. Advanced Technologies and Societal Change. Springer Berlin Heidelberg, S. 123–141. doi: 10.1007/978-3-642-37988-8_9 (siehe S. 3, 11, 19, 153).

— (2017). "A feedback estimation approach for therapeutic facial training". In: *12th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. Accepted. Washington, D.C., USA (siehe S. 10).

Dong, Junyu et al. (2008). "An approach for quantitative evaluation of the degree of facial paralysis based on salient point detection". In: *Intelligent Information Technology Application Workshops (IITAW)*, S. 483–486 (siehe S. 14, 181).

Duda, Richard O., Peter E. Hart und David G. Stork (2001). *Pattern classification*. 2. Aufl. Wiley-Interscience (siehe S. 113, 114, 119).

Dunker, Peter, Stefanie Nowak, André Begau und Cornelia Lanz (2008). "Mood classification for photos and music: A generic multi-modal classification framework and evaluation approach". In: *ACM International Conference on Multimedia Retrieval (MIR)*. Vancouver, Canada, S. 97–104 (siehe S. 11).

Ekman, Paul und Wallace V. Friesen (1976). "Measuring facial movement". In: *Environmental Psychology and Nonverbal Behavior* 1.1, S. 56–75 (siehe S. 115).

Evin, Mark und Sonny Bae. *Jintronix*. Letzter URL-Zugriff: 16.12.2015. url: http://www.jintronix.com (siehe S. 12).

Fang, Tianhong, Xi Zhao, Omar Ocegueda, Shishir K. Shah und Ioannis A Kakadiaris (2011). "3d facial expression recognition: A perspective on promises and challenges". In: *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, S. 603–610 (siehe S. 57).

Flynn, Patrick J und Anil K Jain (1989). "On reliable curvature estimation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 110–116 (siehe S. 93, 98).

Gaber, Amira, Mona F. Taher und Manal Abdel Wahed (2015). "Quantifying facial paralysis using the kinect v2". In: *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, S. 2497–2501 (siehe S. 15, 186).

Gebhard, Arnd, Dietrich Paulus, Bruno Suchy, I. Fucak, S. Wolf und Heinrich Niemann (2001). "Automatische Graduierung von Gesichtsparesen". In: *5. Workshop Bildverarbeitung für die Medizin*. Lübeck, S. 352–356. URL: `http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2001/Gebhard01-AGV.pdf` (siehe S. 14, 180, 183).

Gebhard, Arnd, Dietrich Paulus, Bruno Suchy und Wolf (2000). "A system for diagnosis support of patients with facialis paresis". In: *Künstliche Intelligenz (Kl)* 3/2000. URL: `http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2000/Gebhard00-ASF.pdf` (siehe S. 5, 15, 17, 179).

Graumann, Walther und Dieter Sasse (2004). *CompactLehrbuch Anatomie: in 4 Bänden*. Bd. 4. Schattauer Verlag (siehe S. 174).

Haase, Daniel, Laura Minnigerode, Gerd Fabian Volk, Joachim Denzler und Orlando Guntinas-Lichius (2015). "Automated and objective action coding of facial expressions in patients with acute facial palsy". In: *European Archives of Oto-Rhino-Laryngology* 272.5, S. 1259–1267 (siehe S. 14, 186).

Hadlock, Tessa A. und Luke S. Urban (2012). "Toward a universal, automated facial measurement tool in facial reanimation". In: *Archives of Facial Plastic Surgery* 14.4, S. 277–282 (siehe S. 14, 184).

Harris, Chris und Mike Stephens (1988). "A combined corner and edge detector". In: *Proceedings of the Alvey Vision Conference*, S. 147–152 (siehe S. 42).

Hartley, Richard und Andrew Zisserman (2003). *Multiple view geometry in computer vision*. Cambridge University Press (siehe S. 31, 37).

He, Shu, John J. Soraghan und Brian F. O'Reilly (2007). "Biomedical image sequence analysis with application to automatic quantitative assessment of facial paralysis". In: *Journal on Image and Video Processing* 2007.3 (siehe S. 14, 181).

— (2008). “Supervised Local Linear Embedding (SLLE) for facial paralysis image sequence analysis”. In: *IEEE International Conference on Multimedia and Expo*, S. 49–52 (siehe S. 5, 14, 17, 182).

He, Shu, John J. Soraghan, Brian F. O’Reilly und Dongshan Xing (2009). “Quantitative analysis of facial paralysis using local binary patterns in biomedical videos”. In: *IEEE Transactions on Biomedical Engineering* 56.7, S. 1864–1870 (siehe S. 14, 171, 182).

Herrera, Daniel, Juho Kannala und Janne Heikkilä (2011). “Accurate and practical calibration of a depth and color camera pair”. In: *Computer Analysis of Images and Patterns*, S. 437–445 (siehe S. 187).

House, John W. und Derald E. Brackmann (1985). “Facial nerve grading system.” In: *Official journal of American Academy of Otolaryngology-Head and Neck Surgery* 93.2 (siehe S. 13).

Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin et al. (2003). *A practical guide to support vector classification*. National Taiwan University. URL: `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf` (siehe S. 67, 137).

Huynh, Tri, Rui Min und Jean-Luc Dugelay (2013). “An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data”. In: *Asian Conference on Computer Vision - Workshops*, S. 133–145 (siehe S. 62, 63, 192).

Jähne, Bernd (2002). *Digitale Bildverarbeitung*. 5. Aufl. Springer-Verlag (siehe S. 29, 30).

Jayatilake, Dushyantha, Takashi Isezaki, Anna Gruebler, Youhei Teramoto, Kiyoshi Eguchi und Kenji Suzuki (2012). “A wearable robot mask to support rehabilitation of facial paralysis”. In: *4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, S. 1549–1554 (siehe S. 5, 15, 183).

Khoshelham, Kourosh und Sander Oude Elberink (2012). “Accuracy and resolution of kinect depth data for indoor mapping applications”. In: *Sensors* 12.2, S. 1437–1454 (siehe S. 38–40, 46, 187).

Kihara, Yuta, Guifang Duan, Takeshi Nishida, Naoki Matsushiro und Yen-Wei Chen (2011). “A dynamic facial expression database for quantitative analysis

of facial paralysis". In: *6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*, S. 949–952 (siehe S. 183).

Kim, Hyun Seok, So Young Kim, Young Ho Kim und Kwang Suk Park (2015). "A smartphone-based automatic diagnosis system for facial nerve palsy". In: *Sensors* 15.10, S. 26756–26768 (siehe S. 14, 186).

Kittel, Anita M. (2008). *Myofunktionelle Störungen: Ein Ratgeber für Eltern und erwachsene Betroffene*. 3. Aufl. Schulz-Kirchner Verlag (siehe S. 1, 18, 19).

Kleiss, Ingrid J., Marc H. Hohman, Olivia E. Quatela, Henri A.M. Marres und Tessa A. Hadlock (2013). "Computer-assisted assessment of ocular synkinesis: A comparison of methods". In: *The Laryngoscope* 123.4, S. 879–883 (siehe S. 14, 184).

Lai, Kevin, Liefeng Bo, Xiaofeng Ren und Dieter Fox (2011). "A large-scale hierarchical multi-view rgb-d object dataset". In: *IEEE International Conference on Robotics and Automation (ICRA)*, S. 1817–1824 (siehe S. 49).

Lanz, Cornelia, Joachim Denzler und Horst-Michael Gross (2013a). "Mimikdysfunktionen: Konzeption eines therapiebegleitenden Trainingssystems". In: *Deutscher Ambient-Assisted-Living Kongress (AAL)*. Berlin, Germany, S. 186–195 (siehe S. 9).

— (2013b). "Robust landmark localization for facial therapy applications". In: *European Conference on Technically Assisted Rehabilitation (TAR)*. Berlin, Germany (siehe S. 9, 153, 154).

Lanz, Cornelia, Hanna Lukashevich und Stefanie Nowak (2010a). "Automated classification of film scenes based on film grammar". In: *Workshop Audiovisuelle Medien (WAM)*. Chemnitz, Germany, S. 143–155 (siehe S. 11).

Lanz, Cornelia, Stefanie Nowak und Uwe Kuehhirt (2010b). "Determination of categories for tagging and automated classification of film scenes". In: *European Conference on Interactive TV and Video (EuroITV)*. Tampere, Finland, S. 297–300 (siehe S. 11).

Lanz, Cornelia, Birant Sibel Olgay, Joachim Denzler und Horst-Michael Gross (2013c). "Automated classification of therapeutic face exercises using the kinect". In: *International Conference on Computer Vision Theory and Applications (VISAPP)*. Barcelona, Spain, S. 556–565 (siehe S. 10, 153).

— (2014). “Facial Landmark Localization and Feature Extraction for Therapeutic Face Exercise Classification”. In: *Computer Vision, Imaging and Computer Graphics – Theory and Applications*. Communications in Computer and Information Science. Springer, S. 179–194. DOI: 10.1007/978-3-662-44911-0 (siehe S. 10).

Lemaire, Pierre, Mohsen Ardabilian, Liming Chen und Mohamed Daoudi (2013). “Fully automatic 3D facial expression recognition using differential mean curvature maps and histograms of oriented gradients”. In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, S. 1–7 (siehe S. 62, 63, 87, 109, 191).

Lemaire, Pierre, Boulbaba Ben Amor, Mohsen Ardabilian, Liming Chen und Mohamed Daoudi (2011). “Fully automatic 3D facial expression recognition using a region-based approach”. In: *Proceedings of the Joint ACM Workshop on Human Gesture and Behavior Understanding*, S. 53–58 (siehe S. 60–63, 189).

Li, Chao und Antonio Soares (2011). “Automatic facial expression recognition using 3D faces”. In: *International Journal of Engineering Research & Innovation* 3.1 (siehe S. 59–61, 189).

Li, Xiaoli, Qiuqi Ruan und Yue Ming (2010). “3D facial expression recognition based on basic geometric features”. In: *10th International Conference on Signal Processing (ICSP)*, S. 1366–1369 (siehe S. 59–61, 188).

Lin, Liu Yi (2013). “3D facial model analysis for clinical medicine”. Diss. National University of Singapore (siehe S. 15, 184).

Liu, Li’an, Guanglin Cheng, Junyu Dong, Shengke Wang und Haibin Qu (2010). “Evaluation of facial paralysis degree based on regions”. In: *Workshop on Knowledge Discovery and Data Mining (WKDD)*, S. 514–517 (siehe S. 182).

Lowe, David G. (2004). “Distinctive image features from scale-invariant keypoints”. In: *International Journal of Computer Vision* 60.2, S. 91–110 (siehe S. 92).

Maalej, Ahmed, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava und Stefano Berretti (2011). “Shape analysis of local facial patches for 3D facial expression recognition”. In: *Pattern Recognition* 44.8, S. 1581–1589 (siehe S. 61–63, 189).

Maalej, Ahmed, B. Ben Amor, Mohamed Daoudi, Anuj Srivastava und Stefano Berretti (2010). “Local 3D shape analysis for facial expression recognition”. In:

20th International Conference on Pattern Recognition (ICPR), S. 4129–4132 (siehe S. 61–63, 188).

Matyunin, Sergey, Dmitriy Vatolin, Yury Berdnikov und Michail Smirnov (2011). "Temporal filtering for depth maps generated by kinect depth camera". In: *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, S. 1–4 (siehe S. 49).

Menna, Fabio, Fabio Remondino, Roberto Battisti und Erica Nocerino (2011). "Geometric investigation of a gaming active device". In: *SPIE Optical Metrology*. International Society for Optics und Photonics, 80850G–80850G (siehe S. 37, 187).

Microsoft Corporation. *Developing with Kinect for Windows*. Letzter URL-Zugriff: 28.01.2016. url: `https://dev.windows.com/en-us/kinect/develop` (siehe S. 7).

— *Kinect For Xbox 360*. Letzter URL-Zugriff: 16.01.2016. url: `http://www.xbox.com/de-DE/xbox-360` (siehe S. 12, 22, 37).

Modersohn, Luise und Joachim Denzler (2016). "Facial paresis index prediction by exploiting active appearance models for compact discriminative features". In: *International Conference on Computer Vision Theory and Applications (VISAPP)* (siehe S. 14, 186).

Mpiperis, Iordanis, Sotiris Malassiotis und Michael G. Strintzis (2008). "Bilinear elastically deformable models with application to 3d face and facial expression recognition". In: *8th IEEE International Conference on Automatic Face and Gesture Recognition*, S. 1–8 (siehe S. 57).

Nakamura, Katsuhiko, Naoki Toda, Koichiro Sakamaki, Kenji Kashima und Noriaki Takeda (2003). "Biofeedback rehabilitation for prevention of synkinesis after facial palsy". In: *Otolaryngology–Head and Neck Surgery* 128.4, S. 539–543 (siehe S. 14, 180).

Nintendo of Europe GmbH. *Wii*. Letzter URL-Zugriff: 16.01.2016. url: `http://www.nintendo.de/Wii/Wii-94559.html` (siehe S. 12).

— (2007-2010). *FACE TRAINING - Übungen von Fumiko Inudo*. Letzter URL-Zugriff: 18.12.2015. url: `www.nintendo.de` (siehe S. 17).

Nöth, Elmar et al. (2011). "Automatic evaluation of dysarthric speech and telemedical use in the therapy". In: *The Phonetician* 103.1, S. 75–87. url: `http://www5.`

informatik.uni-erlangen.de/Forschung/Publikationen/2011/Noeth11-AEO.pdf (siehe S. 183).

Okreu, Susanne und Martina Beckers (2006). *Mundmotorik & Fazialisübungen*. Hippocampus Verlag KG (siehe S. 20).

Olgay, Birant Sibel (2012). *Konzepterstellung und Merkmalsuntersuchung für die Entwicklung eines Systems zur Analyse von Gesichtsmotorik-Übungen mittels Tiefeninformation*. Masterarbeit. Technische Universität Ilmenau, Fachgebiet Neuroinformatik und Kognitive Robotik (siehe S. 21, 79, 168, 171).

Park, Byung-Hwa und Se-Young Oh (2015). "Facial expression training system using bilinear shape model". In: *Proceedings of the 3rd International Conference on Human-Agent Interaction*, S. 239–241 (siehe S. 185).

Pears, Nick, Yonghuai Liu und Peter Bunting (2012). *3D imaging, analysis and applications*. Springer (siehe S. 33).

Probst, Rudolf, Gerhard Grevers und Heinrich Iro (2008). *Hals-Nasen-Ohren-Heilkunde*. Georg Thieme Verlag (siehe S. 18, 174).

Rabiu, Habibu, M. Iqbal Saripan, Syamsiah Mashohor und Mohd Hamiruce Marhaban (2012). "3D facial expression recognition using maximum relevance minimum redundancy geometrical features". In: *EURASIP Journal on Advances in Signal Processing* 2012.1, S. 1–8 (siehe S. 59–61, 65, 68–72, 77, 107–109, 112, 170, 190).

Rahmann, S. und H. Burkhardt (2011). *Praktikumsversuch Kamerakalibrierung*. http://lmb.informatik.uni-freiburg.de/lectures/praktika_brox/bvpraktikum/BVAnl_kam_kalib.pdf, letzter Zugriff: 10.03.2014. (Siehe S. 31).

Ramanathan, Subramanian, Ashraf Kassim, Y.V. Venkatesh und Wu Sin Wah (2006). "Human facial expression recognition using a 3D morphable model". In: *IEEE International Conference on Image Processing*. IEEE, S. 661–664 (siehe S. 57).

Reddy, P. Rajashekar, V. Amarnadh und Mekala Bhaskar (2012). "Evaluation of stopping criterion in contour tracing algorithms". In: *International Journal of Computer Science and Information Technologies* 3.3, S. 3888–3894 (siehe S. 53).

Ross, Brenda G, Gaeton Fradet und Julian M. Nedzelski (1996). "Development of a sensitive clinical facial grading system". In: *Otolaryngology–Head and Neck Surgery* 114.3, S. 380–386 (siehe S. 13).

Rusu, Radu Bogdan und Steve Cousins (2011). "3D is here: Point Cloud Library (PCL)". In: *IEEE International Conference on Robotics and Automation (ICRA)*. Letzter URL-Zugriff: 28.01.2016. Shanghai, China. url: `http://pointclouds.org` (siehe S. 7, 64).

Sandbach, Georgia, Stefanos Zafeiriou und Maja Pantic (2012a). "Binary pattern analysis for 3D facial action unit detection". In: (Siehe S. 62, 63, 190).

— (2012b). "Local normal binary patterns for 3D facial action unit detection". In: *19th IEEE International Conference on Image Processing (ICIP)*, S. 1813–1816 (siehe S. 62, 63, 171, 189).

Savran, Arman, Ruben Gur und Ragini Verma (2013). "Automatic detection of emotion valence on faces using consumer depth cameras". In: *IEEE International Conference on Computer Vision Workshops (ICCVW)*, S. 75–82 (siehe S. 62, 63, 191).

Savran, Arman, Bülent Sankur und M Taha Bilge (2010). "Facial action unit detection: 3D versus 2D modality". In: *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, S. 71–78 (siehe S. 63, 188).

Schaffernicht, Erik (2012). "Lernbeiträge im Rahmen einer kognitiven Architektur für die intelligente Prozessführung". Diss. Technische Universität Ilmenau, Fachgebiet Neuroinformatik und Kognitive Robotik (siehe S. 67).

Schmidt, R. und G. Thews (1993). *Physiologie des Menschen*. Springer-Verlag (siehe S. 174).

Schmied, Stefanie und Ulrich Schmied. *Computer Mund Zunge*. Letzter URL-Zugriff: 18.12.2015. url: `http://www.comuzu.de/` (siehe S. 17).

Schreiber, Andreas (2015). *Sprechbegleiter*. Letzter URL-Zugriff: 18.12.2015. Medando UG. url: `http://medando.de/produkte/therapie/sprechbegleiter-2/` (siehe S. 17).

Schwenkreis, Peter (2012). "Periphere Fazialisparese: Eine Blickdiagnose?" In: *Der Allgemeinarzt* 34.2, S. 22–24. url: `http://www.allgemeinarzt-online.de/a/1562679` (siehe S. 18, 19).

Scovanner, Paul, Saad Ali und Mubarak Shah (2007). "A 3-dimensional sift descriptor and its application to action recognition". In: *Proceedings of the 15th International Conference on Multimedia*, S. 357–360 (siehe S. 93).

Smisek, Jan, Michal Jancosek und Tomas Pajdla (2013). "3D with Kinect". In: *Consumer Depth Cameras for Computer Vision*. Springer, S. 3–25 (siehe S. 46, 187).

Sokolova, Marina und Guy Lapalme (2009). "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4, S. 427–437 (siehe S. 68, 154).

Soyel, Hamit und Hasan Demirel (2007). "Facial expression recognition using 3D facial feature distances". In: *Image Analysis and Recognition*. Springer, S. 831–838 (siehe S. 59, 60, 187, 188).

— (2008). "3D facial expression recognition with geometrically localized facial features". In: *23rd International Symposium on Computer and Information Sciences (ISCIS)*, S. 1–4 (siehe S. 59, 60, 188).

Srivastava, Ruchir und Sujoy Roy (2009). "3D facial expression recognition using residues". In: *TENCON 2009-2009 IEEE Region 10 Conference*. IEEE, S. 1–5 (siehe S. 59, 60, 188).

Steuer, Ralf, Jürgen Kurths, Carsten O. Daub, Janko Weise und Joachim Selbig (2002). "The mutual information: detecting and evaluating dependencies between variables". In: *Bioinformatics* 18.suppl 2, S. 231–240 (siehe S. 121).

Sun, Yi und Lijun Yin (2008). "Facial expression recognition based on 3D dynamic range model sequences". In: *European Conference on Computer Vision (ECCV)*, S. 58–71 (siehe S. 57).

Tang, Hao und Thomas S. Huang (2008). "3D facial expression recognition based on automatically selected features". In: *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, S. 1–8 (siehe S. 59, 60, 188).

Tang, Hao und Thomas S Huang (2008). "3D facial expression recognition based on properties of line segments connecting facial feature points". In: *8th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, S. 1–6 (siehe S. 59–61, 188).

Tang, Shuai et al. (2013). "Histogram of oriented normal vectors for object recognition with a depth sensor". In: *Asian Conference on Computer Vision (ACCV)*, S. 525–538 (siehe S. 87).

Tasneem, Tamanna, Atanu Shome und S.K. Alamgir Hossain (2014). "A gaming approach in physical therapy for facial nerve paralysis patient". In: *16th International Conference on Computer and Information Technology (ICCIT)*, S. 345–349 (siehe S. 5, 16, 17, 185).

Taylor, Matthew J.D., Darren McCormick, Rebecca Impson, T. Shawis und Murray Griffin (2011). "Activity promoting gaming systems in exercise and rehabilitation." In: *Journal of Rehabilitation Research and Development* 48.10, S. 1171–1186 (siehe S. 4).

Topi, Mäenpää, Ojala Timo, Pietikäinen Matti und Sariano Maricor (2000). "Robust texture classification by subsets of local binary patterns". In: *15th International Conference on Pattern Recognition*. Bd. 3. IEEE, S. 935–938 (siehe S. 62).

Vretos, Nicholas, Nikos Nikolaidis und Ioannis Pitas (2011). "3D facial expression recognition using Zernike moments on depth images". In: *18th IEEE International Conference on Image Processing (ICIP)*, S. 773–776 (siehe S. 63, 189).

Wachtman, Galen Ss, Jeffrey F. Cohn, Jessie M. VanSwearingen und Ernest K. Manders (2001). "Automated tracking of facial features in patients with facial neuromuscular dysfunction". In: *Plastic and Reconstructive Surgery* 107.5, S. 1124–1133 (siehe S. 180).

Waldeyer, Anton und Anton Mayet (1986). *Anatomie des Menschen. Zweiter Teil: Kopf und Hals, Auge, Ohr, Gehirn, Arm, Brust*. Walter de Gruyter (siehe S. 174).

Wang, Jun, Lijun Yin, Xiaozhou Wei und Yi Sun (2006). "3D facial expression recognition based on primitive surface feature distribution". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Bd. 2, S. 1399–1406 (siehe S. 59–63, 67, 96, 109, 110, 187).

Wang, Peng, Christian Kohler, Fred Barrett, Raquel Gur und R. Verma (2007). "Quantifying facial expression abnormality in schizophrenia by combining 2D and 3D features". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–8 (siehe S. 60–63, 188).

Wang, Shaoyu und Feihu Qi (2005). "Compute aided diagnosis of facial paralysis based on Pface". In: *27th Annual International Conference of the Engineering in Medicine and Biology Society (EMBS)*, S. 4353–4356. doi: 10.1109/IEMBS.2005.1615429 (siehe S. 14, 181).

Wang, Ting, Shu Zhang, Junyu Dong, Li'an Liu und Hui Yu (2015). "Automatic evaluation of the degree of facial nerve paralysis". In: *Multimedia Tools and Applications*, S. 1–16 (siehe S. 14, 186).

Wang, Yiding und Xiaolei Ma (2012). "3D facial expression recognition based on encoded templates". In: *Symposium on Photonics and Optoelectronics (SOPO)*, S. 1–4 (siehe S. 61–63, 190).

Wang, Yiding, Meng Meng und Qingkai Zhen (2013). "Learning encoded facial curvature information for 3D facial emotion recognition". In: *Seventh International Conference on Image and Graphics (ICIG)*, S. 529–532 (siehe S. 62, 63, 109, 192).

Wang, Yong-Xiang, Li-Yun Lo und Min-Chun Hu (2014). "Eat as much as you can: A kinect-based facial rehabilitation game based on mouth and tongue movements". In: *Proceedings of the 22nd ACM International Conference on Multimedia*. Orlando, Florida, USA, S. 743–744. isbn: 978-1-4503-3063-3. doi: 10.1145/2647868.2654887. url: http://doi.acm.org/10.1145/2647868.2654887 (siehe S. 5, 16, 17, 185).

Wickelmaier, Florian (2003). *An introduction to MDS*. Sound Quality Research Unit, Aalborg University, Denmark (siehe S. 120).

Wolowski, Annette (2005). "Fehlregenerationen des Nervus facialis–ein vernachlässigtes Krankheitsbild". Diss. Westfällische Wilhelms-Universität Münster, Medizinische Fakultät (siehe S. 2, 14).

Xue, Mingliang, Ajmal Mian, Wanquan Liu und Ling Li (2014). "Fully automatic 3D facial expression recognition using local depth features". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, S. 1096–1103 (siehe S. 62, 63, 192).

Yin, Lijun, Xiaozhou Wei, Yi Sun, Jun Wang und Matthew J. Rosato (2006). "A 3d facial expression database for facial behavior research". In: *Proceedings of the International Conference on Face and Gesture Recognition*, S. 211–216 (siehe S. 59, 69, 107, 108).

Yu, Yu, Yonghong Song, Yuanlin Zhang und Shu Wen (2013). "A shadow repair approach for kinect depth maps". In: *Asian Conference on Computer Vision (ACCV)*, S. 615–626 (siehe S. 47, 48, 50, 52).

Yun, Tie und Ling Guan (2010). "Human emotion recognition using real 3D visual features from gabor library". In: *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, S. 505–510 (siehe S. 63, 189).

Zeng, Wei, Huibin Li, Liming Chen, J.-M. Morvan und Xianfeng David Gu (2013). "An automatic 3D expression recognition framework based on sparse representation of conformal images". In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, S. 1–8 (siehe S. 60, 63, 192).

Zhai, Meng-yao, Guo-dong Feng und Zhi-qiang Gao (2008). "Facial grading system: Physical and psychological impairments to be considered". In: *Journal of Otology* 3.2, S. 61–67. URL: `http://www.sciencedirect.com/science/article/pii/S167229300850016X` (siehe S. 13).

Zhang, Zhengyou (2000). "A flexible new technique for camera calibration". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.11, S. 1330–1334 (siehe S. 37).

Zhu, Xiangxin und Deva Ramanan (2012). "Face detection, pose estimation, and landmark localization in the wild". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2879–2886 (siehe S. 154).

Index

Erklärung gemäß Anlage 1 der Promotionsordnung

Ich versichere, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Der in dieser Arbeit verwendete Evaluationsdatensatz wurde in Zusammenarbeit mit Birant Sibel Olgay aufgenommen und gelabelt. Im Rahmen ihrer Masterarbeit, welche von mir betreut wurde, entstand zudem eine erste Implementierung des Punktsignaturextraktionsalgorithmus, auf welche die in dieser Arbeit verwendete Implementierung aufbaut.

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer Prüfungsbehörde vorgelegt.

Ich bin darauf hingewiesen worden, dass die Unrichtigkeit der vorstehenden Erklärung als Täuschungsversuch bewertet wird und gemäß § 7 Abs. 10 der Promotionsordnung den Abbruch des Promotionsverfahrens zur Folge hat.

Ilmenau, 03.06.2016 .....

Cornelia Dittmar